PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L 15/02, 15/06

A1

(11) International Publication Number: WO 00/31723

(43) International Publication Date: 2 June 2000 (02.06.00)

(21) International Application Number: PCT/US99/25978

(22) International Filing Date: 3 November 1999 (03.11.99)

(30) Priority Data:
09/200,227 25 November 1998 (25.11.98) US

(71) Applicant: SONY ELECTRONICS, INC. [US/US]; 1 Sony Drive, Park Ridge, NJ 07656 (US).

(72) Inventors: CHEN, Ruxin; Apartment 23, 1600 Petersen Avenue, San Jose, CA 95129 (US). OLORENSHAW, Lex, S.; 267 Morningside Drive, Corte Madera, CA 94925 (US). TANAKA, Miyuki; 1300 Zanker Road, M/S: SJ2D4, San Jose, CA 95134 (US). MENENDEZ-PIDAL, Xavier; 1975 Las Encinas Court, Los Gatos, CA 95032 (US).

(74) Agents: SOBRINO, Maria, E. et al.; Blakely, Sokoloff, Taylor & Zafman, 7th floor, 12400 Wilshire Boulevard, Los Angeles, CA 90025-1026 (US).

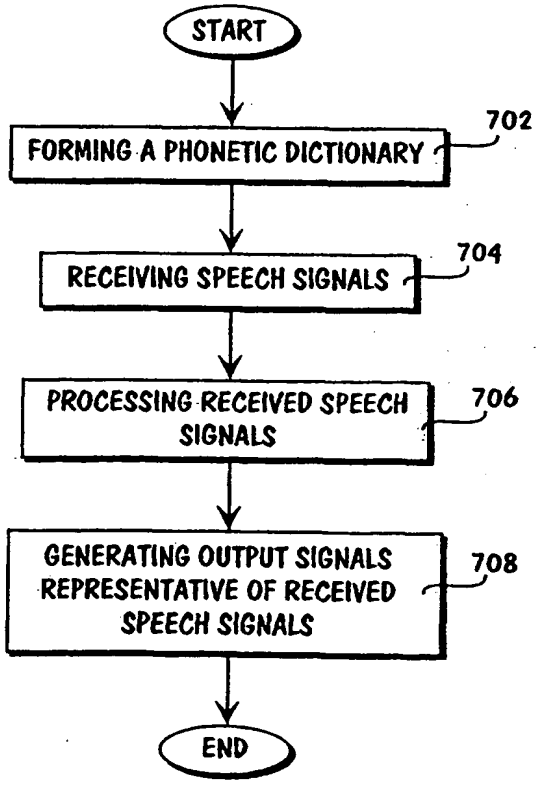
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR VERY LARGE VOCABULARY ISOLATED WORD RECOGNITION IN A PARAMETER SHARING SPEECH RECOGNITION SYSTEM

(57) Abstract

A very large vocabulary isolated word speech recognition system is provided, wherein speech signals are received into a processor. A phonetic dictionary is formed from a baseform dictionary by applying up to four sets of phonological rules to generate phonetic spelling variations for each word. The spelling variations may account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation. The received speech signals are processed using a speech recognition system comprising the phonetic dictionary and at least one phoneme set. In one embodiment, the phoneme set comprises a single phoneme to account for stop closures and glottal stops. Moreover, the phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel. Furthermore, the speech recognition system is produced by generating a number of phoneme models, some of which are shared among a number of phonemes. Output signals are generated that are representative of the received speech signals.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD AND APPARATUS FOR VERY LARGE VOCABULARY ISOLATED WORD RECOGNITION IN A PARAMETER SHARING SPEECH RECOGNITION SYSTEM

RELATED APPLICATION

The present application is a continuation-in-part of United States Patent Application Number 08/953,026, filed October 16, 1997.

FIELD OF THE INVENTION

This invention relates to speech or voice recognition systems. More particularly, this invention relates to a speech recognition system based on a parameter sharing phoneme model.

BACKGROUND OF THE INVENTION

The broad goal of speech recognition technology is to create devices that can receive spoken information and act appropriately upon that information. In order to maximize benefit and universal applicability, speech recognition systems (SRSs) should be capable of recognizing isolated speech, and should be able to recognize multiple speakers with possibly diverse accents, speaking styles, and different vocabularies and grammatical tendencies. Effective SRSs should also be able to recognize poorly articulated speech, and should have the ability to recognize speech in noisy environments.

Acoustic models of sub-word sized speech units form the backbone of virtually all SRSs. Many systems use phonemes to define the dictionary, but some SRSs use allophones. The best recognition performance is typically obtained when acoustic models are generated for the sub-word units conditioned on their context; such models are called context-dependent sub-word models. When the chosen sub-word unit is the phoneme, the context-dependent modeling can capture allophonic

variation and coarticulation. In the case of phones, context-dependent modeling only attempts to capture the effects of coarticulation.

Once a speaker has formed a thought to be communicated to the listener, they construct a phrase or sentence by choosing from a collection of sounds, or phonemes. The basic theoretical unit for describing how speech conveys linguistic meaning is called a phoneme. As such, the phonemes of a language comprise a minimal theoretical set of units that are sufficient to convey all meaning in the language; this is to be compared with the actual sounds that are produced in speaking, which speech scientists call allophones. For American English, there are approximately 50 phonemes which are made up of vowels, semivowels, diphthongs, and consonants. Each phoneme can be considered to be a code that consists of a unique set of articulatory gestures. If speakers could exactly and consistently produce these phoneme sounds, speech would amount to a stream of discrete codes. However, because of many different factors including, for example, accents, gender, and coarticulatory effects, every phoneme has a variety of acoustic manifestations in the course of flowing speech. Thus, from an acoustical point of view, the phoneme actually represents a class of sounds that convey the same meaning.

The most abstract problem involved in speech recognition is enabling the speech recognition system with the appropriate language constraints. Whether phones, phonemes, syllables, or words are viewed as the basic unit of speech, language, or linguistic, constraints are generally concerned with how these fundamental units may be concatenated, in what order, in what context, and with what intended meaning. For example, if a speaker is asked to speak a phoneme in isolation, the phoneme will be clearly identifiable in the acoustic waveform. However, when spoken in context, phoneme boundaries become difficult to label because of the physical properties of the speech articulators. Since the vocal tract articulators consist of human tissue, their positioning from one phoneme to the next is executed by movement of muscles that control articulator movement. As such, there is a period of transition between phonemes that can modify the manner in

which a phoneme is produced. Therefore, associated with each phoneme is a collection of allophones, or variations on phones, that represent acoustic variations of the basic phoneme unit. Allophones represent the permissible freedom allowed within a particular language in producing a phoneme, and this flexibility is dependent on the phoneme as well as on the phoneme position within an utterance.

Prior art SRSs can recognize phonemes uttered by a particular speaker. A speaker-dependent SRS uses the utterances of a single speaker to learn the models, or parameters, that characterize the SRS's internal model of the speech process. The SRS is then used specifically for recognizing the speech of its trainer. Accordingly, the speaker-dependent SRS will yield relatively high recognition results compared with a speaker-independent SRS. Prior art SRSs also perform speaker-independent recognition. The speaker-independent SRS is trained by multiple speakers and used to recognize many speakers who may be outside of the training population. Although more accurate, the disadvantage of a speaker-dependent SRS is the need to retrain the system each time it is to be used with a new speaker.

At present, the most popular approach in speech recognition is statistical learning, and the most successful statistical learning technique is the hidden Markov model (HMM). The HMMs are capable of robust and succinct modeling of speech, and efficient maximum-likelihood algorithms exist for HMM training and recognition. To date, HMMs have been successfully applied to the following constrained tasks: speaker-dependent recognition of isolated words, continuous speech, and phones; small-vocabulary speaker-independent recognition of isolated words; and speaker-independent phone recognition in large vocabulary continuous and isolated word recognition.

The HMMs provide a sound basis for modeling both the interspeaker and intraspeaker variability of natural speech. However, to accurately model the distributions of real speech spectra, it is necessary to have complex output distributions. For example, continuous density HMM systems require multiple Gaussian mixture components to achieve good performance. Furthermore, context-dependent triphones are required to deal with contextual effects such as

coarticulation. Thus, a speaker-independent continuous speech HMM system will generally contain a large number of context-dependent models, each of which contains a large number of parameters. Unfortunately, the ability to arbitrarily increase model complexity is limited by the limited amount of training data and the statistical confidence of this data. Thus, the key problem to be faced when building a HMM-based continuous speech recognizer is maintaining the balance between model complexity, the corresponding processor requirements, and the available training data, and finding the best method by which to estimate the model parameters.

Traditional methods of dealing with this problem tend to be model-based. For example, for discrete and tied-mixture systems it is common to interpolate between triphones, biphones and monophones. One prior art technique of speaker-independent phone recognition generates a model based on multiple codebooks of linear predictive coding-derived parameters for a number of phones and then applies co-occurrence smoothing to determine the similarity between every pair of codewords from all phones, smoothing the individual distributions accordingly. However, a speaker-independent phone model is unstable because in actual speech the context depends on the preceding and the following phone or maybe even the phone two positions before or after; thus, each different context of a phone requires a different model which increases the speech recognition system memory requirements as well as decreasing system accuracy, efficiency, and speed.

In an attempt to avoid the need for smoothing, both stochastic decision trees and maximum *a posteriori* estimation approaches have been proposed. Another prior art speech recognition method produces a context-dependent Gaussian mixture HMM in which acoustic phone states are merged and then any cluster with insufficient training data is merged with its nearest neighbor. There also exists a prior art speech recognition system in which phones are clustered depending on their phonetic context into left and right contexts. However, one of the limitations of the prior art model-based approaches is that the left and right contexts cannot be

treated independently and since the distribution of training examples between left and right contexts will rarely be equal, this leads to a suboptimal use of the data.

In addition to the HMM, another approach available in speech recognition is the knowledge engineering approach. Knowledge engineering techniques integrate human knowledge about acoustics and phonetics into a phone recognizer, which produces a sequence or a lattice of phones from speech signals. While hidden Markov learning places learning entirely in the training algorithm, the knowledge engineering approach attempts to explicitly program human knowledge about acoustic/phonetic events into the speech recognition system. Whereas an HMM-based search is data driven, a knowledge engineering search is typically heuristically guided. Currently, knowledge engineering approaches have exhibited difficulty in integrating higher level knowledge sources with the phonetic decoder as a result of decoder complexity. Consequently, there is a requirement for a speech recognition system that combines knowledge engineering in an interchangeable way with stochastic methods including HMMs comprising phoneme models to produce and use a model for very large vocabulary isolated word speech recognition that reduces memory requirements of the SRS while maximizing the use of available training data to reduce the error in parameter estimation and optimize the training result.

SUMMARY OF THE INVENTION

A method and an apparatus for very large vocabulary isolated word recognition in a parameter sharing speech recognition system are provided. The speech recognition system comprises a speaker-independent, English isolated, very large vocabulary system, and may be used in systems comprising car navigation systems, information retrieval systems, and database query systems. According to one aspect of the invention, speech signals are received into a processor. A phonetic dictionary is formed from a baseform dictionary by applying phonological rules to generate phonetic spelling variations for each word. The phonological rules comprise vowel variation rules and consonant variation rules. The baseform

dictionary may be formed from a base set of phonetic symbols. The spelling variations may account for acoustic variations comprising dialect, phonological processes of the language, acoustic-phonetic processes of the language, and overpronunciation. The received speech signals are processed using a speech recognition system comprising the phonetic dictionary and at least one phoneme set. In one embodiment, the phoneme set comprises a single phoneme to account for stop closures and glottal stops. Moreover, the phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel. Furthermore, the speech recognition system is produced by generating a number of phoneme models, portions of which are shared among a number of phonemes. Output signals are generated that are representative of the received speech signals.

The formation of the phonetic dictionary of an embodiment comprises the application of a number of sets of phonological rules. A first set of phonological rules provides a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for every word in the baseform dictionary. This allows each entry in a phonetic dictionary (which is not necessarily constructed for use with the SRS) to be represented in a form which is now usable by the SRS. A second set of phonological rules provides for the inclusion of closure phonemes before stops and glottal stop phonemes before word-initial vowels. A third set of phonological rules provides for the formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation. A fourth set of phonological rules provides for the formation of phonetic dictionary entries by analyzing at least one transcription of a training data set to account for pronunciation variations. A fifth set of rules provides for the mapping of a base set of phone symbols to a particular set of phone symbols used by a phone set of the speech recognition system.

These and other features, aspects, and advantages of the present invention will be apparent from the accompanying drawings and from the detailed description and appended claims which follow.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements and in which:

Figure 1 shows a speech signal as a one-dimensional waveform.

Figure 2 shows one trained phonetic HMM topology for phone /d/.

Figure 3 shows a multiple-state left-to-right HMM phoneme model of an embodiment of the present invention.

Figure 4 shows a speech recognition system of an embodiment of the present invention.

Figure 5 is a computer system hosting the speech recognition system (SRS) of an embodiment of the present invention.

Figure 6 is the computer system memory hosting the speech recognition system of an embodiment of the present invention.

Figure 7 is a flowchart for recognizing speech using a speech recognition system of an embodiment of the present invention.

Figure 8 shows the five phone sets used in an embodiment of the present invention.

Figure 9 shows a speech waveform for the word "item" along with phonetic transcriptions corresponding to the phone set P1 and phone sets P2-P5 of an embodiment of the present invention.

Figure 10 shows the isolated word recognition results using five different phone sets and two different dictionaries with the speech recognition system of an embodiment of the present invention.

Figure 11 is a flowchart for producing the parameter sharing HMM used in an embodiment of the present invention.

Figure 12 shows the structure, or topology, of a parameter sharing HMM in an embodiment of the present invention.

Figure 13 is another depiction of the structure of a parameter sharing HMM in an embodiment of the present invention.

Figure 14 is a flowchart for the method of generating the shared phoneme models in an embodiment of the present invention.

Figure 15 shows phoneme model sharing between two triphone models having a common biphone in an embodiment of the present invention.

Figure 16 shows state sharing between two triphone phoneme models in an embodiment of the present invention.

Figure 17 shows PDF sharing between two phoneme model states 1 and 2 in an embodiment of the present invention.

Figure 18 shows Gaussian PSDF sharing between two PDFs 1802 and 1804 in a continuous-time observation HMM of an embodiment of the present invention.

Figure 19 is a flowchart for producing the parameter sharing HMM of an embodiment of the present invention using top-down reevaluation.

Figure 20 is a flowchart for producing the parameter sharing HMM of an embodiment of the present invention using bottom-up reevaluation.

Figure 21 shows an HMM structure following further sharing using a bottom-up approach in an embodiment of the present invention.

Figure 22 shows an HMM structure following further sharing using a top-down approach in an embodiment of the present invention.

Figure 23 shows the 110K isolated word recognition accuracies that are experimentally observed using HMM structures M1-M5 of an embodiment of the present invention.

Figure 24 shows the isolated word recognition accuracy results using the four versions of the 110K dictionaries D1-D4 and the M5 HMMs of 2583 states with the P5 phone set of an embodiment of the present invention.

Figure 25 shows the isolated word recognition accuracy results using the four versions of the 110K dictionaries D1-D4 and the M6 HMMs of 2655 states with the P4 phone set of an embodiment of the present invention.

Figure 26 shows the isolated word recognition accuracy results using dictionaries of different sizes and types, without training transcription, and the M5 HMMs of an embodiment of the present invention.

Figure 27 shows the isolated word recognition accuracy results using dictionaries of different sizes and types, with training transcription, and the M5 HMMs of an embodiment of the present invention.

Figure 28 shows the isolated word recognition accuracy results using P4 dictionaries of different sizes and types, with training transcription, and the M6 HMMs of an embodiment of the present invention.

Figure 29 shows a summary of improvements on the very large vocabulary English isolated word recognition using a speech recognition system of an embodiment of the present invention.

Figure 30a and b shows the results obtained with a speech recognition system of an embodiment of the present invention compared to isolated word and continuous speech recognition system results reported in the literature.

DETAILED DESCRIPTION

A method and an apparatus for very large vocabulary isolated word recognition in a parameter sharing speech recognition system are provided. The method and apparatus described herein may also be used in pattern recognition systems. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be evident, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention. It is noted that preliminary experiments with the method and apparatus provided herein show significant speech recognition improvements when compared to typical prior art speech recognition systems.

Figure 1 shows a speech signal 100 as a one-dimensional waveform. The speech signal 100 corresponding to words in a continuously spoken utterance can be segmented into words. The speech signal 100 is comprised of the words "whatever" 102 and "there" 104 in a continuously spoken sentence. The speech

signal 100 can be labeled with a sequence of phonemes wherein each word 102 and 104 comprises one or more continuous phonemes 110-126. The word "whatever" 102 substantially comprises the phonemes "w" 110, "ah" 112, "dx" 114, "eh" 116, "v" 118, and "er" 120. The word "the" 104 substantially comprises the phonemes "dh" 122, "eh" 124, and "r" 126.

An HMM is a stochastic finite state automaton, or a type of abstract machine used to model a speech utterance. The utterance modeled by an HMM of one embodiment may be, but is not limited to, a word, a subword unit like a phoneme, or a complete sentence or paragraph. Using the HMM, a speech utterance is reduced to a string of features, or observations, because these features represent the information that is "observed" from the incoming speech utterance. Therefore, an HMM which is associated with a particular phoneme or other utterance is a finite state machine capable of generating observation strings. An HMM is more likely to produce observation strings that would be observed from real utterances of its associated phoneme.

The HMM is used in two phases of speech recognition. In the training phase, the HMM is trained as to the statistical makeup of the observation strings for its dedicated phoneme. In the recognition phase the HMM receives as input a given incoming observation string, and it is imagined that one of the existing HMMs produced the observation string. The phoneme associated with the HMM of highest likelihood is declared to be the recognized word.

Figure 2 shows one trained phonetic HMM 100 topology for phone /d/. The structure, or topology, of the HMM is determined by its allowable state transitions. This model represents the phone /d/ using seven states 201-207 and twelve transitions 220-231. The HMM generates observation sequences by jumping from state to state and emitting an observation with each jump. The HMM generally used for modeling acoustic signals emits an observation upon arrival at each successive state. At each observation time, corresponding to the times at which observations are extracted from the speech utterances to be recognized, a state transition is assumed to occur in the model. The likelihood of these transitions is

governed by the state transition probabilities. These state transition probabilities appear as labels on the transitions 220-231, or arcs connecting the states 201-207. The sequence of states that occurs enroute to generating a given observation sequence defines the first of two random processes associated with an HMM, the matrix of state transition probabilities or the state transition matrix. The state transition matrix taken together with an initial state probability vector completely specifies the probability of residing in any state at any time.

A left-to-right HMM process is used to model the speech waveform in the SRS of one embodiment. **Figure 3** shows a multiple-state left-to-right HMM phoneme model 300 of an embodiment of the present invention. A series of HMMs corresponds to a series of phonemes. Therefore, HMM phoneme model 300 will be preceded and followed by other HMM phoneme models, the other HMM phoneme models being similar to HMM phoneme model 300 in one embodiment. The HMM phoneme model 300 is comprised of three states 302-306. States are used to represent identifiable acoustic phenomena. Therefore, the number of states is often chosen to correspond to the expected number of such phenomena in the utterance. When HMMs are used to model phonemes, three states are typically used: one state for the onset transition, one state for the steady-state portion of the phone, and one state for the exiting transition. The three-state HMM 300 is used to model a context-dependent phoneme in the context of its left and right phones. The observations are emitted from each state, and the observations emitted from each state form a distribution that can be formulated as a probability distribution b_s, where s refers to a state in the HMM. Each state transition 312-316 is associated with a state transition probability $a_{s,j}$ which denotes the probability of a transition using the arc j of state s. For example, suppose there are B types of observations, and b_{si} denotes the distribution of state s and type i, then

$$b_s = \sum_i b_{si}, i = 1 \ldots B$$

Observation vectors may contain, but are not limited to, many features that are used as observations. The most frequently used general features of speech used as observations include linear prediction (LP) parameters, cepstral parameters, and related quantities derived from the voltage levels of the speech signal, the power contained in the speech signal, and the energy present in a particular frequency band. These are frequently supplemented by short-term time differences that capture the dynamics of the signal, as well as energy measures such as the short-term energy and differenced energy. For example, in a typical application of the HMM, the incoming speech signal is sampled at a particular frequency, for example 8 kHz or higher, and analyzed on frames of a specified number of points having a specified overlap. These sample times become observation times. Multiple LP coefficients are then computed for each frame. These LP coefficients are converted to a multiple number of cepstral coefficients. In order to add dynamic information, a number of differenced cepstral coefficients are also included in the vector. Furthermore, a short-term energy measure and a differenced energy measure are included for each frame.

In a speech recognition system, an observation sequence generally may be modeled as either a discrete-time stochastic process or a continuous-time stochastic process. When the observation sequence is modeled as a discrete-time stochastic process, the generation of particular observations upon entering a state is governed by the observation probability sub-distribution for that state. The discrete observation HMM produces a finite set of discrete observations. The naturally occurring observation vectors are quantized into one of a permissible set using vector quantization methods. Prior to training any of the HMMs for the individual utterance, a set of continuous observation vectors from a large corpus of speech is used to derive a codebook. If there are Q possible vectors, or observations, in the codebook, then it is sufficient to assign an observation a single integer, q, where

$$1 \leq q \leq Q$$

Subsequently, any observation vector used for either training or recognition is quantized using this codebook. For the discrete observation HMM in one embodiment, the distribution of state s and type i is a one dimensional array described by

$$b_{si} = b_{si}[q] = \sum_k b_{sik}[q], q = 1 \ldots Q$$

where each scalar $b_{si}[q]$ denotes the probability of observing the vector quantized symbol q for state s; $b_{sik}[q]$ denotes the sub-distribution that is comprised of $b_{si}[q]$. The Q in the equation denotes the total number of q. The sub-distribution b_{sik} for the discrete HMM allows for better compression of the discrete HMM parameters and for better sharing of the structure between the discrete HMM and the continuous HMM.

In the more general case in which the observation sequence is modeled as a continuous-time stochastic process, the observations are vector-valued and correspond to the unquantized vectors of the aforementioned features drawn from the speech. Therefore, the formal description of the HMM contains a multivariate PDF characterizing the distribution of observations within each state. For the continuous observation HMM in one embodiment, the distribution of state s and type i is

$$b_{si} = b_{si}(o) = \sum_k b_{sik}(o),$$

where

$$b_{sik}(o) = \frac{c_{sik}}{\sqrt{|v_{sik}|}} \times \exp\left(-0.5 \times \frac{(o - m_{sik})^2}{v_{sik}} \right)$$

A diagonal Gaussian mixture is used to represent the probability of the continuous observation vector o for state s. The variable c_{sik} is the weight for mixture k of state s, type i. Similarly, the variable m_{sik} is the mean for the Gaussian of mixture k. The variable v_{sik} is the variance for mixture k.

As the discrete-time observation HMM is restricted to the production of a finite set of discrete observations, the quantized observation PDF for a state takes the form of impulses on a real line instead of a characteristic distribution. In contrast, the continuous-time observation HMM observation description contains a multivariate PDF characterizing the distribution of observations within each state. Consequently, prior art systems typically use either discrete-time or continuous-time observation models. One embodiment of the present invention integrates, or unifies, discrete observation modeling and continuous observation modeling by generating shared characteristic PDFs for discrete-time observations from continuous observation PDFs. This shared characteristic PDF is then divided into simple segments, or a simple probability sub-distribution function (PSDF), where the PSDFs are shared by both the continuous HMM and the discrete HMM.

As previously discussed, the naturally occurring observation vectors, or cepstrums, for each phoneme frame sample are quantized into a finite number so that the discrete observation HMM produces a finite set of quantized discrete observations for a phoneme. In one embodiment, the finite discrete observations for each cepstrum of each frame of each training data sample are plotted from the vector quantized discrete observations, and regions are established on the plot using statistical techniques known in the art. Centroids are established for each region, and quantized vectors of the incoming speech signals are assigned to the region which minimizes the distance between the quantized vector and the centroid. A PDF is then generated for each frame sample of a phoneme from the distribution of the quantized vectors in the corresponding regions using known statistical techniques.

Figure 4 shows a speech recognition system 400 of an embodiment of the present invention. An input device 402 is coupled to the SRS 400 and inputs a voice

signal 401 into the SRS 400 by converting the voice signal 401 into an electrical signal representative of the voice signal 401. A signal sampler 404 coupled to the input device 402 samples the signal at a particular frequency, the sampling frequency determined using techniques known in the art. A coefficient generator and converter 406 coupled to the signal sampler 404 computes cepstrum or LP coefficients or other speech features and converts these to cepstral coefficients. A signal segmenter 408 coupled to the coefficient generator and converter 406 segments the electrical signal representative of a voice signal into phonemes or phones or words, but is not so limited. A model device 410 coupled to receive the output of the signal segmenter 408 hosts a parameter sharing HMM that is used to model the speech utterance 401. The model device 410 of an embodiment may comprise acoustic models 414 and a language model 416 to perform a hypothesis search, but the embodiment is not so limited. These models are trained in a supervised paradigm as to the statistical makeup of appropriate exemplars, or observation strings. The model device 410 provides output signals 412 representative of the received speech signals 401. The SRS comprising components 402-410 may be hosted on a processor, but is not so limited. For an alternate embodiment, the model device 410 may comprise some combination of hardware and software that is hosted on a different processor from SRS components 402-408. For another alternate embodiment, a number of model devices, each comprising a different model, may be hosted on a number of different processors. Another alternate embodiment has multiple processors hosting a single model. For still another embodiment, a number of different model devices may be hosted on a single processor.

The SRS comprising components 402-416 may be hosted on a processor, but is not so limited. For an alternate embodiment, the model device 410 may comprise some combination of hardware, firmware, and software that is hosted on a different processor from SRS components 402-408. For another alternate embodiment, a number of model devices, each comprising a different acoustic model or a language model, may be hosted on a number of different processors. Another alternate

embodiment has multiple processors hosting the acoustic models and the language model. For still another embodiment, a number of different model devices may be hosted on a single processor.

Figure 5 is a computer system 500 hosting the speech recognition system (SRS) of an embodiment of the present invention. The computer system 500 comprises, but is not limited to, a system bus 501 that allows for communication among a processor 502, a digital signal processor 508, a memory 504, and a mass storage device 507. The system bus 501 is also coupled to receive inputs from a keyboard 522, a pointing device 523, and a speech signal input device 525, but is not so limited. The system bus 501 provides outputs to a display device 521 and a hard copy device 524, but is not so limited.

Figure 6 is the computer system memory 610 hosting the speech recognition system of an embodiment of the present invention. An input device 602 provides speech signals to a digitizer and bus interface 604. The digitizer 604, or feature extractor, samples and digitizes the speech signals for further processing. The digitizer and bus interface 604 allows for storage of the digitized speech signals in the speech input data memory component 618 of memory 610 via the system bus 608. The digitized speech signals are processed by a digital processor 606 using algorithms and data stored in the components 612-622 of the memory 610. As discussed herein, the algorithms and data that are used in processing the speech signals are stored in components of the memory 610 comprising, but not limited to, a hidden Markov model (HMM) training and recognition processing computer program 612, a viterbi processing computer program code and storage 614, a preprocessing computer program code and storage 616, language model memory 620, and acoustic model memory 622.

Figure 7 is a flowchart for recognizing speech using a speech recognition system of an embodiment of the present invention. The speech recognition system comprises a speaker-independent, English isolated, very large vocabulary system, and may be used in systems comprising car navigation systems, information retrieval systems, and database query systems, but the embodiment is not so

limited. Operation begins at step 702, at which a phonetic dictionary is formed from a baseform dictionary by applying phonological rules to generate phonetic spelling variations for each word. The phonetic dictionary of an embodiment comprises approximately 110,000 words, but the embodiment is not so limited. The phonological rules comprise vowel variation rules and consonant variation rules, but the embodiment is not so limited. The baseform dictionary may be formed from a plurality of phonological symbols, or a base set of phonetic symbols, but the embodiment is not so limited. The spelling variations may account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation, but the embodiment is not so limited. Speech signals are received into at least one processor, at step 704. The received speech signals are processed, at step 706, using a speech recognition system comprising the phonetic dictionary and at least one phoneme set. In one embodiment, the phoneme set comprises a single phoneme to account for stop closures and glottal stops, but the embodiment is not so limited. Moreover, the phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel. The phoneme set may be used in at least one recognizer dictionary, but the embodiment is not so limited. Furthermore, the received speech signals may be processed using a speech recognition system produced by generating a number of phoneme models, some of which are shared among a number of phonemes. Signals are generated that are representative of the received speech signals, at step 708.

The formation of the phonetic dictionary of an embodiment may comprise the application of four sets of phonological rules, but the embodiment is not so limited. A first set of phonological rules are applied to entries of the baseform dictionary. The first set of phonological rules provides a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for every word in the baseform dictionary, but the embodiment is not so limited.

A second set of phonological rules may be applied to entries of the baseform dictionary following application of the first set of phonological rules, but the

embodiment is not so limited. The second set of phonological rules provides for the inclusion of closure phonemes before stops and glottal stop phonemes before word-initial vowels.

A third set of phonological rules may be applied to entries of the baseform dictionary following application of the first and second sets of phonological rules, but the embodiment is not so limited. The third set of phonological rules provides for the formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

A fourth set of phonological rules may be applied to entries of the baseform dictionary following application of the first, second, or third sets of phonological rules, but the embodiment is not so limited. The fourth set of phonological rules provides for the formation of phonetic dictionary entries by analyzing at least one transcription of a training data set to account for pronunciation variations.

A fifth set of rules may be applied to entries of the baseform dictionary following application of the first, second, third, and fourth set of phonological rules, but the embodiment is not so limited. This set of rules provides for the mapping of a set of phone symbols to a particular set of phone symbols used by a phone set of an embodiment of the present invention, but the embodiment is not so limited.

Figure 8 shows the five phone sets used in an embodiment of the present invention. The phone set P1 is the phone set used in the Carnegie Mellon University (CMU) 110K dictionary, a dictionary containing mostly words having a single phonetic spelling. The phone sets P2-P5 are designed to take into account acoustic-phonetic variations that are distinct enough and frequent enough to consider modeling separately. Phone set P2 introduces a reduced mid-central unstressed vowel /ax/. For Phone set P2, /ah0/ in the CMU dictionary is converted to /ax/. Phone set P3 introduces a new phone symbol /clq/ that accounts for stop closures and glottal stops in the English language; therefore, for phone set P3, /clq/ is inserted as an option before all stops and word-initial vowels. Phone set P4 introduces a reduced high-central unstressed vowel /ix/. For phone set P4, /ih0/ in the CMU dictionary is converted to /ix/. Phone set P5 uses an

extra glottal stop /q/, separate from the closure. In addition, the closure is separated into three cases: voiced closure /vcl/, voiceless closure /cl/, and epenthetic closure /epi/. Other symbols added in P5 are for syllabic "n" (/en/) syllabic "l" (/el/), and flapped "t" or "d" (/dx/). **Figure 9** shows a speech waveform 900 for the word "item" along with phonetic transcriptions corresponding to the phone set P1 and phone sets P2-P5 of an embodiment of the present invention. The transcription indicated as "→" 902 indicates a segment of speech attributed to the next phone symbol.

As discussed herein, the speech recognition system of an embodiment is a speaker-independent, very large vocabulary, English isolated word recognition system comprising a dictionary comprising approximately 110,000 entries, but the embodiment is not so limited. Anticipated uses of the speech recognition system comprise vehicle navigation systems, information retrieval systems, and database query systems.

In order to develop a robust general purpose recognition system for very large vocabulary isolated words, the system is built on the phoneme concept. Five different phone sets were used to model the English phonemes, and experiments were conducted to determine the phone set that produces the best results. The speech corpus presented herein is designed to be suitable for the training and evaluation of very large vocabulary isolated word recognition systems, but the embodiment is not so limited. The speech corpus was recorded in a sound-treated room at Sony US Research Labs. Approximately 50 hours of isolated word data were recorded. The corpus comprises words that account for the triphones that appear in the 2000 most frequent English words. This word list was composed by taking the common words among the 5000 most frequent word lists from the Brown Corpus {H. Kucera, et al. (1967); Brown Corpus; Computational analysis of present-day American English Providence, Brown University Press}, the British National Corpus {Adam Kilgarriff, et al. (1996); British National Corpus; ftp://ftp.itri.bton.ac.uk/pub/bnc}, and the Switchboard Corpus {Bill Byrne, et al. (1996); WS96 Switchboard Data Resources (1996);

ftp://homer.clsp.jhu.edu/pub/swbdWS96}. To extend the training vocabulary, a set of less common words were added from the above three sources and a set of randomly chosen words from a 50K dictionary. The total number of unique recorded words for training was 11,584, but the embodiment is not so limited. The performance evaluation of the speech recognition system of one embodiment is based on a separate set of 20 independent test speakers (9633 word tokens of 1783 unique words). Test words came from frequent English words that are present in the training vocabulary.

Initially, monophone HMMs having 16 mixtures per state were used to evaluate the recognition performance of the CMU dictionary with different phone sets. Subsequent experimental models were built using triphone context dependent continuous HMMs having 3 left-to-right states, wherein each state comprises 16 Gaussian mixtures.

Figure 10 shows the isolated word recognition results using five different phone sets and two different dictionaries with the speech recognition system of an embodiment of the present invention. The recognition accuracies for the top three dictionary candidates are shown. One of the recognition dictionaries comprises the Sony 5K most frequent English words; this is a composite version of the 5000 most frequent words of English, based on common words in the aforementioned three top 5K lists. It is noted that the stop closures and reduced vowels of an embodiment of the present invention result in a significant improvement in the isolated word recognition accuracy, especially when using triphone models, but the embodiment is not so limited. Furthermore, different phone sets have different impacts on the results depending on the dictionary selected. By tuning the phone set of the CMU dictionary, the recognition accuracy may be improved from 67 percent to 90 percent.

Figure 11 is a flowchart for producing the parameter sharing HMM used in an embodiment of the present invention. The parameter sharing HMM used by the model device 410 of one embodiment of the present invention is based on a statistical learning approach that comprises multiple phoneme models. The

parameter sharing HMM of one embodiment utilizes sharing among a number of model levels and sharing within each model level. Production of the HMM begins at step 1102, at which multiple context-dependent phoneme models are generated in which some of the phoneme models are shared among multiple phonemes. The structure, or topology, of the generated HMM is based, at least in part, on the amount of training data available. Once generated, the HMM is trained, at step 1104, using a particular library of training data as selected by the system designer. Training the HMM to represent a phoneme amounts to finding a procedure for estimating an appropriate state transition matrix and observation PDFs for each state.

Following the training of the HMM, a number of phoneme model states are generated, at step 1106. These phoneme model states are representative of the phoneme models and may include shared states generated using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing. At step 1108, a number of phoneme model probability distribution functions (PDFs) are generated. These PDFs are representative of the phoneme model states, and may include shared PDFs generated using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing. Following generation of the PDFs, a number of shared probability sub-distribution functions (PSDFs) are generated, at step 1110. These PSDFs are representative of the phoneme model PDFs, and may include shared PSDFs generated using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing. At step 1112, the shared phoneme models are evaluated for further sharing in light of the shared phoneme model states, PDFs, and PSDFs. This reevaluation of the sharing hierarchy for further sharing may be accomplished using a top-down approach or a bottom-up approach or a combination top-down and bottom-up approach as will be discussed herein.

Figure 12 shows the structure, or topology, of a parameter sharing HMM in an embodiment of the present invention. This parameter sharing HMM is hosted

by the model device 410 of the SRS, in one embodiment. The structure of the parameter sharing HMM is comprised of four levels 1291-1294 and, in this example, parameter sharing is used among all levels 1291-1294. The first level 1291 comprises the phoneme models 1200-1203 of the HMM. The second level 1292 comprises the phoneme model states 1204-1209 used to represent the phoneme models 1200-1203. In this embodiment, each phoneme model comprises three phoneme model states, but is not so limited. The third level 1293 comprises the phoneme model PDFs 1210-1216 used to represent the phoneme model states 1204-1209. In this embodiment, each phoneme model state comprises two PDFs, but is not so limited. The fourth level 1294 comprises the PSDFs 1217-1227 used to represent the phoneme model PDFs. In this embodiment, each phoneme model PDF comprises three PSDFs, but is not so limited.

An example of parameter sharing shown in **Figure 12** is the sharing of phoneme model state 1206 by phoneme models 1200 and 1202. Another example of parameter sharing is the sharing of phoneme model PDF 1210 by phoneme model states 1204 and 1205. Still another example of parameter sharing is the sharing of phoneme model PSDF 1219 by phoneme model PDFs 1210, 1211, and 1212.

Moreover, parameters can be shared within levels. For example, two states 1205 and 1208 of level 1292 included in a model 1203 may be statistically similar resulting in the generation of one phoneme model state 1250 to represent these two states. As another example, phoneme model PSDFs 1223 and 1225 of level 1293 may be statistically similar resulting in the generation of one phoneme model PSDF 1260 to represent these two PSDFs.

Figure 13 is another depiction of the structure of a parameter sharing HMM in an embodiment of the present invention. Phoneme models 1301 and 1303 are shared within level 1391 to generate shared phoneme model 1321. Shared phoneme model 1321 shares phoneme model states 1306, 1308, and 1309 at level 1392. Phoneme model state 1320 is statistically representative of the three states comprised in phoneme model 1300 so that phoneme model 1300 at level 1391 shares with phoneme model state 1320 at level 1392. Phoneme model PDFs 1311 and 1313

are shared within level 1393 to generate shared phoneme model PDF 1322. Shared phoneme model PDF 1322 at level 1393 is shared by phoneme model states 1305, 1306, 1307, 1309, and 1310 at level 1392. Shared phoneme model PDF 1322 at level 1393 shares phoneme model PSDFs 1315 and 1316. Phoneme model PSDF 1318 is statistically representative of phoneme model PDF 1314 so that phoneme model PDF 1314 at level 1393 shares with phoneme model PDSF 1318 at level 1394.

Figure 14 is a flowchart for the method of generating the shared phoneme models in an embodiment of the present invention. The method of generating the shared phoneme models uses knowledge engineering techniques that integrate human knowledge about acoustics and phonetics into phoneme model generation to generate a hierarchy of sharing. The generation of the shared phoneme model does not require the actual training library data; instead the number of data, or frame, samples in the training library for each phoneme model is used. Using this data, at step 1402, any triphone model having a number of trained frames available in the training library that exceeds a prespecified threshold is retained as a separate phoneme model. The threshold may be a predetermined level of statistical significance, but is not so limited. Furthermore, any phoneme model deemed to be important to a system designer may be retained.

After removing the models retained at step 1402 from consideration, and after removing the frames used in step 1402 to generate the retained models, a shared phoneme model is generated, at step 1404, to represent each of the groups of triphone phoneme models for which the number of trained frames available in the training library having a common biphone exceed the prespecified threshold. The common biphone may comprise either the center context in combination with the right context of the triphone model, or the center context in combination with the left context of the triphone model. In the aforementioned sharing, the amount of training data, or number of frames, available for each model was used to determine the sharing structure.

Figure 15 shows phoneme model sharing between two triphone models having a common biphone in an embodiment of the present invention. For the

context dependent HMM, the model comprising "w-ah+dx" means the phoneme "ah" having the left context phoneme "w" and the right context phoneme "dx". In this sharing example, triphones 1502, 1504, and 1506 share the same center context "ah" and right context "dx". Taken together, the center context and the right context comprise a biphone 1508. Therefore, one triphone phoneme model is generated having the statistical properties of the center context "ah" and the right context "dx". This phoneme model is used anywhere in the HMM that the phoneme model for any of the components 1502-1506 is required. Using knowledge engineering, it may be possible to generate a triphone phoneme model comprising the common biphone and having a left context with statistical properties that approximate the statistical properties of many of the component 1502-1506 left contexts.

After removing the models retained and generated at steps 1402 and 1404 from consideration, and after removing the frames used in steps 1402 and 1404 to generate the retained models, a shared phoneme model is generated, at step 1406, to represent each of the groups of triphone phoneme models for which the number of trained frames available in the training library having an equivalent effect on a phonemic context exceed the prespecified threshold. This step is where a large amount of knowledge engineering is used in evaluating the "equivalent effect". The equivalent effect on a phonemic context for a center context may be an equivalent sound, but is not so limited. The equivalent effect on a phonemic context for a left and a right context may be an equivalent impact on the center context by the left and the right context, but is not so limited.

After removing the models retained and generated at steps 1402-1406 from consideration, and after removing the frames used in steps 1402-1406 to generate the retained models, a shared phoneme model is generated, at step 1408, to represent each of the groups of triphone phoneme models having the same center context.

After removing the models retained and generated at steps 1402-1408 from consideration, and after removing the frames used in steps 1402-1408 to generate the retained models, a shared phoneme model is generated, at step 1410, based on

context duration data. Moreover, a shared triphone model may be generated to represent a group of phonemes wherein each context of the shared triphone model comprises statistical properties of a group of context phonemes.

As previously discussed, after the shared phoneme models are generated, the models are trained using the training library data. Following this training of the shared phoneme models, a multiple number of shared PSDFs are generated from the trained phoneme models. The actual training library data, or frames, is used to generate these PSDFs.

The generation of the shared PSDFs begins by generating a number of shared states from the states comprising each of the shared phoneme HMMs. The states represent segments of a phoneme of a speech signal. As previously discussed, three states are used in one embodiment: one state for the onset transition portion of the phoneme speech signal, one state for the steady-state portion of the phoneme speech signal, and one state for the exiting transition portion of the phoneme speech signal. The shared states have the same state transition probability and the same observation distribution. The shared states are generated by using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to combine states.

Figure 16 shows state sharing between two triphone phoneme models in an embodiment of the present invention. In this sharing example, triphones 1610 and 1612 share states 1602 and 1604. Triphone 1610 is comprised of phoneme "ah" with left context "ao", and right context "iy"; triphone 1610 is represented by states 1602, 1602, and 1604, respectively. Triphone 1612 is comprised of phoneme "ah" with left context "ah" and right context "iy"; triphone 1612 is represented by states 1604, 1606, and 1604, respectively.

The generation of the shared PSDFs continues by generating a number of shared phoneme model PDFs from the PDFs comprising each of the shared phoneme model states. In one embodiment, each state can have up to four PDFs, but is not so limited. The PDFs are generated, as previously discussed, from LP parameters, cepstral parameters, and related quantities derived from the voltage

levels, power, and energy contained in a speech signal. For example, four often-used PDFs are generated from signal power plots, cepstral coefficients, differenced cepstral coefficients, and differenced cepstral coefficients. The shared PDFs are generated by using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to combine PDFs.

Figure 17 shows PDF sharing between two phoneme model states 1 and 2 in an embodiment of the present invention. Each of the rectangles 1702-1706 represents one distribution b_{si}. In this sharing example, states 1 and 2 share PDFs 1702, 1704, and 1706. State 1 comprises PDFs 1702 and 1706. State 2 comprises PDFs 1702 and 1704.

The generation of the shared PSDFs continues by generating a number of shared phoneme model PSDFs from the sub-distributions comprising each of the shared phoneme model PDFs. For continuous-time observation HMMs, the mixture of Gaussian distributions $b_{si}(o)$ comprise the shared Gaussian PDFs. For discrete-time observation HMMs, the PDFs $b_{si}[q]$ comprise the shared discrete sub-distributions.

Figure 18 shows Gaussian PSDF sharing between two PDFs 1802 and 1804 in a continuous-time observation HMM of an embodiment of the present invention. Probability distribution 1802 shares Gaussian PSDFs 1806 and 1810. Probability distribution 1804 shares Gaussian PSDFs 1808 and 1810.

Following the generation of a sharing hierarchy for the HMM, and in response to the multiple number of shared PSDFs generated from the trained phoneme models, the sharing hierarchy comprising the shared phoneme models, states, PDFs, and PSDFs is evaluated for further sharing. This reevaluation of the sharing hierarchy for further sharing may be accomplished using a top-down approach or a bottom-up approach.

Figure 19 is a flowchart for producing the parameter sharing HMM of an embodiment of the present invention using top-down reevaluation. The steps 1902-1910 of this top-down approach are the same as steps 1102-1110 of **Figure 11** so that

this top-down approach repeats the steps used when initially generating the shared parameter HMM with the steps having the same ordering. However, if reevaluation of the generated phoneme models is required at step 1912, then operation continues at step 1902 where steps 1902-1910 are now reevaluated in light of the additional information provided by the previous hierarchical sharing. This approach may be repeated as many times as the system designer determines necessary.

Figure 20 is a flowchart for producing the parameter sharing HMM of an embodiment of the present invention using bottom-up reevaluation. When using the bottom-up reevaluation approach, operation begins at step 2001 at which a determination is made that the phoneme models are to be evaluated for further sharing as no sharing has been implemented because no model has been generated. Steps 2002-2010 are the same as steps 1102-1110 of **Figure 11** for initially generating the shared parameter HMM. However, if reevaluation of the generated phoneme models is required at step 2012, then operation continues by proceeding backwards sequentially from step 2010 through step 2001 whereby the bottom-up approach repeats the steps used when initially generating the shared parameter HMM, the steps of **Figure 11**, except the steps are performed in the reverse order. This approach may be repeated as many times as the system designer determines necessary so that at step 2001, if further reevalution is determined to be necessary, then steps 2002 through 2012 are repeated. Using the bottom-up approach, a shared phoneme model PDF is generated to replace any PDFs that share all PSDFs. When all PSDFs are not shared, shared PDFs may still be generated by using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to combine PDFs that are similar.

Following reevaluation of PDF sharing, a shared phoneme model state is generated to replace any states that share all PDFs. When all PDFs are not shared, shared states may still be generated by using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to combine states that are similar.

Following reevaluation of state sharing, a shared phoneme model is generated to replace any models that share all states. When all states are not shared, shared models may still be generated by using a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to combine models that are similar. Either the top-down or bottom-up approach, or some combination thereof, may be repeated as many times as determined to be necessary by the system designer.

With reference to **Figure 12**, some examples of the reevaluation of the sharing hierarchy comprising the shared phoneme models, states, PDFs, and PSDFs for further sharing are provided. The parameter sharing referred to herein uses a combination of statistical techniques known in the art and knowledge engineering in the area of acoustics processing to provide combined models, states, PDFs, and PSDFs. **Figure 21** shows an HMM structure following further sharing using a bottom-up approach in an embodiment of the present invention. Beginning at the bottom, or PSDF, level it is noted that PSDFs 2117, 2118, and 2119 are shared by PDFs 2110 and 2111. Therefore, using parameter sharing , a single PDF 2199 may be generated to be used in the place of PDFs 2110 and 2111. At the PDF level, PDFs 2110 and 2111 are shared by phoneme model states 2104 and 2105. Therefore, using parameter sharing, a single state 2198 may be generated to be used in the place of states 2104 and 2105.

Figure 22 shows an HMM structure following further sharing using a top-down approach in an embodiment of the present invention. Beginning at the phoneme model level, it is noted that models 2201 and 2203 share phoneme model states 2205, 2207, and 2209. Therefore, using parameter sharing, a single phoneme model 2296 may be generated to be used in the place of models 2201 and 2203. At the phoneme model state level, states 2207 and 2208 share PDFs 2212 and 2213. Therefore, a single state 2297 may be generated to be used in the place of states 2207 and 2208. At the PDF level, PDFs 2210 and 2211 share PSDFs 2217, 2218, and 2219. Therefore, a single PDF 2298 may be generated to be used in the place of PDFs 2210

and 2211. Thus, sharing may occur between adjacent levels so that the structure of any particular level shares structures at lower levels of the HMM structure.

Moreover, in the HMM structure of one embodiment sharing may occur at any level between the models, states, PDFs and PSDFs of that level. Thus, for example, PSDFs 2226 and 2227, if statistically similar, may be shared between PDF 2216 to generate a single phoneme PDF 2299. The phoneme PDF 2299, in this case, would also represent a single phoneme PSDF 2299.

The speech recognition system of an embodiment of the present invention supports arbitrary sharing of HMMs, HMM states, the mixture-probability distribution (MPDTR), and the sub-probability distribution (SPDTR), wherein the MPDTR and SPDTR are introduced to support a unified sharing structure for both continuous HMMs and discrete HMMs. Using phone set P5, as discussed herein, monophone model M0, left-biphone models M1 and M2, right-biphone model M3, and shared state triphone models M4 and M5 were created. The maximum number of mixtures used per state is 16 for all HMMs except M1, which has 4 mixtures per state, but the embodiment is not so limited.

Figure 23 shows the 110K isolated word recognition accuracies that are experimentally observed using HMM structures M1-M5 of an embodiment of the present invention. It is noted that HMM model structures have a significant impact on isolated word recognition, and that for very large vocabulary English isolated word recognition, triphone modeling with a parameter sharing structure provides the best results. The monophone HMMs might be suitable for recognition using a smaller vocabulary size, but the 53.4 percent top3 recognition accuracy would not typically be acceptable for any actual application. By improving the HMM modeling structure, the system improved the recognition performance from 45.7 percent to 80.7 percent. The difference in recognition accuracies for M4 and M5 demonstrates the importance of an HMM sharing structure in addition to full triphone modeling.

Results of an evaluation using the improved phonetic dictionary also indicated that the speech recognition system of the present invention showed

significant improvements over the performance of typical systems. The isolated word recognizer of the present invention uses no high level linguistic knowledge, such as word bigrams or trigrams, but the embodiment is not so limited. The only context used in the mapping of sub-word (acoustic-phonetic) units to words is the phonetic dictionary, but the embodiment is not so limited. Therefore, the speech dictionary of the present invention described the phonetic spellings as accurately and completely as possible.

As discussed herein, the CMU dictionary D1 is the base dictionary with minor alterations applied via a first set of phonological rules to make a usable version of the dictionary for the SRS of an embodiment of the present invention. A modified dictionary D2, as compared to D1, provides for optional closure phones before stops and optional glottal stop phones before word-initial vowels via the second set of phonological rules. Another modified dictionary, D3, as compared to D2, has some entries from a base dictionary which may use a different base entry for some words which may provide a better base for application of subsequent sets of phonological rules. Primarily, D3 comprises additional phonetic spelling variations based on a set of phonological rules. A fourth version of the dictionary, D4, is derived from D3 with the addition of transcriptions generated as a result of analyzing a training data set. These four dictionaries D1-D4 are remapped accordingly for use with phone sets P4 and P5 via the fifth set of rules, but the embodiment is not so limited.

It is noted that all 1783 unique testing words appear in the training data vocabulary, which comprises 11,584 unique words. Therefore, the inclusion of training transcriptions provides a better phonetic transcription for the test words. Since the test words are from the top 5K English words, this word set comprises many words that are short and phonetically confusing. For example, "coach", "coast", "coat", "code", "coke", and "cold" are included in the test set.

Figure 24 shows the isolated word recognition accuracy results using the four versions of the 110K dictionaries D1-D4 and the M5 HMMs of 2583 states with the P5 phone set of an embodiment of the present invention. It is noted that

significant improvement of the large vocabulary isolated word recognition accuracy is realized by modifying the CMU dictionary. Specifically, the D2 dictionary provides an approximately 23 percent improvement over the D1 dictionary simply by introducing optional closures. The reason for this improvement may be due to the fact that the D2 dictionary accounts for more detailed acoustic-phonetic variations. The improved performance using the D3 and D4 dictionaries shows the significant impact of a phonetic dictionary comprising pronunciation variation. Furthermore, it is noted that the D4 dictionary does not directly include the transcriptions of the test data, but the embodiment is not so limited.

Figure 25 shows the isolated word recognition accuracy results using the four versions of the 110K dictionaries D1-D4 and the M6 HMMs of 2655 states with the P4 phone set of an embodiment of the present invention. It is noted that these results are improved, using fewer states, over the isolated word recognition accuracy results using the four versions of the 110K dictionaries D1-D4 and the M5 HMMs of 2583 states with the P5 phone set of an embodiment of the present invention. As P5 is a superset of P4, better recognition accuracy would likely be expected with more detailed transcription. However, the compactness of P4 compared to P5 likely makes it easier for the speech recognition system of an embodiment to obtain a good HMM sharing structure. Furthermore, the differences in these two results demonstrate the importance of a robust and compact HMM structure like that used in an embodiment of the present invention.

Improvements in very large isolated word recognition make the speech recognition system of an embodiment of the present invention more robust to any vocabulary. Therefore, several experiments were conducted to evaluate the resulting system on other dictionaries. In experiments conducted for the CMU, Sony 50K, and Sony 5K dictionaries, the test tokens are the same as discussed herein. In experiments for the 5K city name dictionary, there are 6125 test tokens from 16 independent speakers, wherein the test tokens comprise 75 unique city names and 25 command words. The actual city name dictionary contains 4927 city names plus 25 commands for car navigation applications.

Figure 26 shows the isolated word recognition accuracy results using dictionaries of different sizes and types, without training transcription, and the M5 HMMs of an embodiment of the present invention. **Figure 27** shows the isolated word recognition accuracy results using dictionaries of different sizes and types, with training transcription, and the M5 HMMs of an embodiment of the present invention. Again, it is noted that the detailed and accurate dictionary phone transcriptions have a significant influence on isolated word recognition, as the recognition accuracies with dictionaries comprising training transcriptions consistently improved about three percent as compared to recognition accuracies with dictionaries not containing training transcriptions.

Figure 28 shows the isolated word recognition accuracy results using P4 dictionaries of different sizes and types, with training transcription, and the M6 HMMs of an embodiment of the present invention. As in the evaluations using the CMU dictionary, the M6 HMM provides better recognition accuracies for all four dictionaries, demonstrating that the HMMs of an embodiment of the present invention are not particularly tuned to any individual vocabulary set.

Figure 29 shows a summary of improvements on the very large vocabulary English isolated word recognition using a speech recognition system of an embodiment of the present invention. These results are sorted in terms of top3 recognition accuracy. It is noted that each improvement in phone set, models, and dictionary spellings helps in achieving an improvement in recognition accuracy for 110K very large vocabulary isolated word recognition from approximately 45.7 percent to approximately 94.6 percent.

Figure 30a and b shows the results obtained with a speech recognition system of an embodiment of the present invention compared to isolated word recognition system results reported in the literature. It is noted that most isolated word recognition results reported in the literature deal with vocabularies of several hundred to several thousand words. In order to include some very large vocabulary recognition results, some continuous speech recognition results are also included.

Thus, a method and apparatus for a very large vocabulary isolated word recognition in a parameter sharing speech recognition system have been provided. Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the invention as set forth in the claims. Accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense.

CLAIMS

What is claimed is:

1. A method for recognizing speech comprising the steps of:

receiving (704) speech signals into a processor;

forming (702) at least one phonetic dictionary from a baseform dictionary by applying at least one set of phonological rules to generate phonetic spelling variations for each word, wherein the spelling variations account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation;

processing (706) the received speech signals using a speech recognition system comprising the at least one phonetic dictionary and at least one phoneme set, the at least one phoneme set comprising a phoneme to account for stop closures and glottal stops; and

generating (708) signals representative of the received speech signals.

2. The method of claim 1, further comprising the step of processing the received speech signals using a speech recognition system produced by generating a plurality of phoneme models, and at least one of the plurality of phoneme models are shared among a plurality of phonemes.

3. The method of claim 1, further comprising the step of forming the baseform dictionary from a plurality of phonetic symbols.

4. The method of claim 1, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a first set of phonological rules to entries of the baseform dictionary, wherein the first set of phonological rules provides a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for at least one entry of the baseform dictionary.

5. The method of claim 1, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a second set of phonological rules to entries of the baseform dictionary, wherein the second set of phonological rules provides at least one closure phoneme before stops and at least one glottal stop phoneme before word-initial vowels.

6. The method of claim 5, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a third set of phonological rules to entries of the baseform dictionary, wherein the third set of phonological rules provides for formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

7. The method of claim 1, wherein the step of forming the at least one phonetic dictionary comprises the steps of:

forming a fourth set of phonological rules from at least one transcription of a training data set;

applying the fourth set of phonological rules, wherein the fourth set of phonological rules accounts for pronunciation variations.

8. The method of claim 1, further comprising the step of remapping the at least one phonetic dictionary for use with another of the at least one phoneme sets.

9. The method of claim 1, wherein the at least one set of phonological rules comprise vowel variation rules and consonant variation rules.

10. The method of claim 1, wherein the at least one phoneme set is used in at least one recognizer dictionary.

11. The method of claim 1, wherein the at least one phonetic dictionary comprises a Carnegie Mellon University (CMU) dictionary.

12. The method of claim 1, wherein the at least one phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel.

13. The method of claim 1, wherein the speech recognition system is based on a statistical learning approach comprising a hidden Markov model (HMM), the speech recognition system comprising triphone context dependent continuous HMMs comprising three left-to-right states, wherein each state has sixteen Gaussian mixtures.

14. The method of claim 1, wherein the speech recognition system is a speaker-independent, English isolated, very large vocabulary system, wherein the at least one phonetic dictionary comprises approximately 110,000 words.

15. The method of claim 1, wherein the processor comprises a car navigation system, an information retrieval system, and a database query system.

16. The method of claim 2, wherein the speech recognition system is produced by:

training (1104) the plurality of phoneme models;

generating (1110) a plurality of shared probability sub-distribution functions from the trained plurality of phoneme models; and

evaluating (1112) the plurality of phoneme models for further sharing in response to the plurality of shared probability sub-distribution functions.

17. The method of claim 2, wherein a plurality of phoneme models are generated by:

retaining (1402) as a separate phoneme model a triphone phoneme model for which a number of trained frames exceeds a threshold;

generating (1404) at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having a common biphone exceeds the threshold;

generating (1406) at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having an equivalent effect on a phonemic context exceeds the threshold; and

generating (1408) at least one shared phoneme model to represent a plurality of triphone phoneme models having the same center context.

18. The method of claim 17, wherein at least one shared phoneme model is generated comprising at least one context, the at least one context having statistical properties representative of a plurality of context phonemes.

19. The method of claim 16, the plurality of shared probability sub-distribution functions generated by:

generating (1906) a plurality of phoneme model states, wherein at least one of the plurality of states are shared among the plurality of phoneme models;

generating (1908)a plurality of phoneme model probability distribution functions, wherein at least one of the plurality of probability distribution functions are shared among the plurality of states; and

generating (1910) a plurality of phoneme model probability sub-distribution functions, wherein at least one of the plurality of probability sub-distribution functions are shared among the plurality of phoneme model probability distribution functions.

20. The method of claim 19, the plurality of shared probability sub-distribution functions generated by:

generating (1906) a plurality of phoneme model states, wherein at least one of the plurality of phoneme model states are shared among the plurality of phoneme model states;

generating (1908) a plurality of phoneme model probability distribution functions, wherein at least one of the plurality of probability distribution functions are shared among the plurality of probability distribution functions; and

generating (1910) a plurality of phoneme model probability sub-distribution functions, wherein at least one of the plurality of probability sub-distribution functions are shared among the plurality of phoneme model probability sub-distribution functions.

21. The method of claim 2, wherein sharing occurs among a plurality of levels of a speech recognition model, wherein sharing occurs within at least one level of a speech recognition model.

22. The method of claim 16, wherein the plurality of phoneme models are evaluated for further sharing by:

generating (2010) a shared phoneme model probability distribution function to replace a plurality of probability distribution functions when each of the plurality of probability distribution functions has common probability sub-distribution functions;

generating (2008) a shared phoneme model state to replace a plurality of states when each of the plurality of states has common phoneme model probability distribution functions; and

generating (2006) a shared phoneme model to replace a plurality of models when each of the plurality of models has common phoneme model states.

23. An apparatus for speech recognition comprising:

a processor;

an input (402) coupled to the processor, the input capable of receiving speech signals, the processor configured to recognize the received speech signals using a speech recognition system (400) comprising at least one phoneme set and at least one phonetic dictionary, the at least one phoneme set comprising a phoneme to account for stop closures and glottal stops, the at least one phonetic dictionary formed from a baseform dictionary by applying at least one set of phonological rules to generate phonetic spelling variations for each word, wherein the spelling variations account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation; and

an output (412) coupled to the processor, the output capable of providing a signal representative of the received speech signal.

24. The apparatus of claim 23, wherein the speech recognition system is produced by generating and training a plurality of phoneme models, wherein at least one of the plurality of phoneme models are shared among a plurality of phonemes.

25. The apparatus of claim 23, wherein the at least one phonetic dictionary is formed by applying a first set of phonological rules to provide a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for at least one entry of the baseform dictionary.

26. The apparatus of claim 23, wherein the at least one phonetic dictionary is formed by applying a second set of phonological rules to provide at least one closure phoneme before stops and at least one glottal stop phoneme before word-initial vowels.

27. The apparatus of claim 26, wherein the at least one phonetic dictionary is formed by applying a third set of phonological rules to provide for

formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

28. The apparatus of claim 23, wherein the at least one phonetic dictionary is formed by applying a fourth set of phonological rules generated from at least one transcription of a training data set, wherein the fourth set of phonological rules accounts for pronunciation variations.

29. The apparatus of claim 23, wherein the at least one phonetic dictionary is remapped for use with another of the at least one phoneme sets.

30. The apparatus of claim 23, wherein the speech recognition system is based on a statistical learning approach comprising a hidden Markov model (HMM), wherein the baseform dictionary is formed from a plurality of phonological symbols, wherein the at least one phoneme set is used in at least one recognizer dictionary, wherein the at least one phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel.

31. The apparatus of claim 24, wherein the speech recognition system is produced by:

generating a plurality of phoneme model states, wherein at least one of the plurality of states are shared among the plurality of phoneme models;

generating a plurality of phoneme model probability distribution functions, wherein at least one of the plurality of probability distribution functions are shared among the plurality of states;

generating a plurality of phoneme model probability sub-distribution functions, wherein at least one of the plurality of probability sub-distribution functions are shared among the plurality of phoneme model probability distribution functions; and

evaluating the plurality of phoneme models for further sharing in response to the plurality of shared probability sub-distribution functions.

32. The apparatus of claim 24, wherein sharing occurs among a plurality of levels of a speech recognition model, and wherein sharing occurs within at least one level of a speech recognition model.

33. The apparatus of claim 24, wherein the plurality of phoneme models are generated by:

retaining as a separate phoneme model a triphone phoneme model for which a number of trained frames exceeds a threshold;

generating at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having a common biphone exceeds the threshold;

generating at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having an equivalent effect on a phonemic context exceeds the threshold; and

generating at least one shared phoneme model to represent a plurality of triphone phoneme models having the same center context.

34. The apparatus of claim 31, wherein the plurality of phoneme models are evaluated for further sharing by:

generating a shared phoneme model probability distribution function to replace a plurality of probability distribution functions when each of the plurality of probability distribution functions has common probability sub-distribution functions;

generating a shared phoneme model state to replace a plurality of states when each of the plurality of states has common phoneme model probability distribution functions; and

generating a shared phoneme model to replace a plurality of models when each of the plurality of models has common phoneme model states.

35. A speech recognition process comprising at least one phoneme set, at least one phonetic dictionary, and a statistical learning technique that uses a model, the at least one phonetic dictionary formed from a baseform dictionary by applying at least one set of phonological rules to generate phonetic spelling variations for each word, wherein the spelling variations account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

36. The speech recognition process of claim 35, wherein the baseform dictionary is formed from a plurality of phonetic symbols, wherein the at least one phoneme set comprises a phoneme to account for stop closures and glottal stops and a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel.

37. The speech recognition process of claim 35, wherein the at least one set of phonological rules comprises a first set of phonological rules that provides a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for at least one entry of the baseform dictionary.

38. The speech recognition process of claim 35, wherein the at least one set of phonological rules comprises a second set of phonological rules that provides at least one closure phoneme before stops and at least one glottal stop phoneme before word-initial vowels.

39. The speech recognition process of claim 35, wherein the at least one set of phonological rules comprises a third set of phonological rules that provides

for formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

40. The speech recognition process of claim 35, wherein the at least one set of phonological rules comprises a fourth set of phonological rules generated from at least one transcription of a training data set, wherein the fourth set of phonological rules accounts for pronunciation variations.

41. The speech recognition process of claim 35, wherein the at least one set of phonological rules comprises a fifth set of phonological rules for remapping the at least one phonetic dictionary for use with another of the at least one phoneme sets.

42. The speech recognition process of claim 35, wherein the model is produced by:

generating and training a plurality of phoneme models, wherein at least one of the plurality of phoneme models are shared among a plurality of phonemes;

generating a plurality of shared probability sub-distribution functions from the trained plurality of phoneme models; and

evaluating the plurality of phoneme models for further sharing in response to the plurality of shared probability sub-distribution functions.

43. The speech recognition process of claim 42, wherein sharing occurs among a plurality of levels of a speech recognition model, and wherein sharing occurs within at least one level of a speech recognition model.

44. The speech recognition process of claim 42, wherein the plurality of phoneme models are context dependent hidden Markov models, wherein the plurality of phoneme models integrate discrete observation modeling and continuous observation modeling.

45. The speech recognition process of claim 42, wherein the plurality of phoneme models are generated by:

retaining as a separate phoneme model a triphone phoneme model for which a number of trained frames exceeds a threshold;

generating at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having a common biphone exceeds the threshold;

generating at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having an equivalent effect on a phonemic context exceeds the threshold; and

generating at least one shared phoneme model to represent a plurality of triphone phoneme models having the same center context.

46. The speech recognition process of claim 42, wherein the plurality of shared probability sub-distribution functions are generated by:

generating a plurality of phoneme model states, wherein at least one of the plurality of states are shared among the plurality of phoneme models;

generating a plurality of phoneme model probability distribution functions, wherein at least one of the plurality of probability distribution functions are shared among the plurality of states; and

generating a plurality of phoneme model probability sub-distribution functions, wherein at least one of the plurality of probability sub-distribution functions are shared among the plurality of phoneme model probability distribution functions.

47. The speech recognition process of claim 42, wherein the plurality of phoneme models are evaluated for further sharing by:

generating a shared phoneme model probability distribution function to replace a plurality of probability distribution functions when each of the plurality of

probability distribution functions has common probability sub-distribution functions;

generating a shared phoneme model state to replace a plurality of states when each of the plurality of states has common phoneme model probability distribution functions; and

generating a shared phoneme model to replace a plurality of models when each of the plurality of models has common phoneme model states.

48. A computer readable medium containing executable instructions which, when executed in a processing system, causes the system to perform the steps of a method for recognizing speech, the method comprising the steps of:

receiving (704) speech signals into a processor;

forming (702) at least one phonetic dictionary from a baseform dictionary by applying at least one set of phonological rules to generate phonetic spelling variations for each word, wherein the spelling variations account for acoustic variations comprising dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation;

processing (706) the received speech signals using a speech recognition system comprising the at least one phonetic dictionary and at least one phoneme set, the at least one phoneme set comprising a phoneme to account for stop closures and glottal stops; and

generating (708) signals representative of the received speech signals.

49. The computer readable medium of claim 48, wherein the method further comprises the steps of:

processing the received speech signals using a speech recognition system comprising a plurality of phoneme models, wherein at least one of the plurality of phoneme models are shared among a plurality of phonemes; and

forming the baseform dictionary from a plurality of phonological symbols.

50. The computer readable medium of claim 48, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a first set of phonological rules to entries of the baseform dictionary, wherein the first set of phonological rules provides a mapping to a base acoustic-phonetic level comprising at least one phonetic spelling variation for at least one entry of the baseform dictionary.

51. The computer readable medium of claim 48, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a second set of phonological rules to entries of the baseform dictionary, wherein the second set of phonological rules provides at least one closure phoneme before stops and at least one glottal stop phoneme before word-initial vowels.

52. The computer readable medium of claim 48, wherein the step of forming the at least one phonetic dictionary comprises the step of applying a third set of phonological rules to entries of the baseform dictionary, wherein the third set of phonological rules provides for formation of phonetic spelling variations for dialect, phonological processes of language, acoustic-phonetic processes of language, and overpronunciation.

53. The computer readable medium of claim 44, wherein the step of forming the at least one phonetic dictionary comprises the steps of:

forming a fourth set of phonological rules from at least one transcription of a training data set;

applying the fourth set of phonological rules, wherein the fourth set of phonological rules accounts for pronunciation variations.

54. The computer readable medium of claim 48, further comprising the step of remapping the at least one phonetic dictionary for use with another of the at least one phoneme sets.

55. The computer readable medium of claim 48, wherein the at least one phoneme set is used in at least one recognizer dictionary, wherein the at least one phoneme set comprises a reduced mid-central unstressed vowel and a reduced high-central unstressed vowel.

56. The computer readable medium of claim 49, wherein the speech recognition system is produced by:

generating (1906) a plurality of phoneme model states, wherein at least one of the plurality of states are shared among the plurality of phoneme models;

generating (1908) a plurality of phoneme model probability distribution functions, wherein at least one of the plurality of probability distribution functions are shared among the plurality of states;

generating (1910) a plurality of phoneme model probability sub-distribution functions, wherein at least one of the plurality of probability sub-distribution functions are shared among the plurality of phoneme model probability distribution functions; and

evaluating the plurality of phoneme models for further sharing in response to the plurality of shared probability sub-distribution functions.

57. The computer readable medium of claim 49, wherein sharing occurs among a plurality of levels of a speech recognition model, and wherein sharing occurs within at least one level of a speech recognition model.

58. The computer readable medium of claim 49, wherein the plurality of phoneme models are generated by:

retaining (1402) as a separate phoneme model a triphone phoneme model for which a number of trained frames exceeds a threshold;

generating (1404) at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having a common biphone exceeds the threshold;

generating (1406) at least one shared phoneme model to represent a plurality of triphone phoneme models for which the number of trained frames having an equivalent effect on a phonemic context exceeds the threshold; and

generating (1408) at least one shared phoneme model to represent a plurality of triphone phoneme models having the same center context.

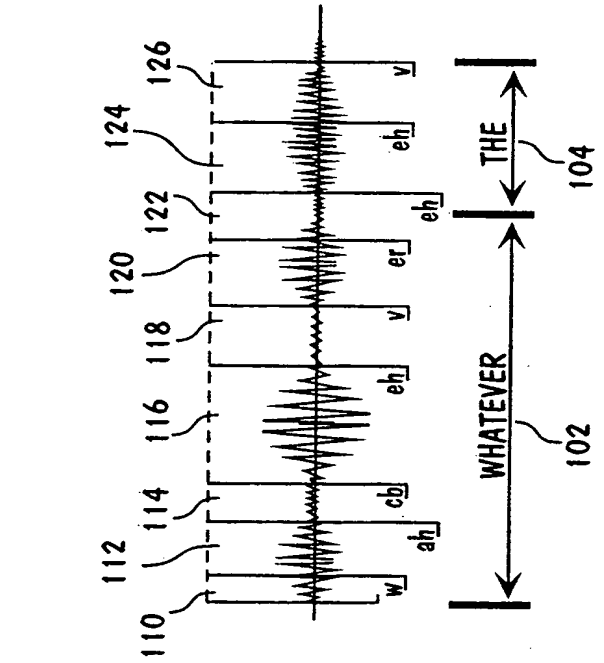

100
110
112
114
116
118
120
122
124
126
w
ah
cb
eh
er
y
eh
eh
y
WHATEVER
102
THE
104

FIG. 1

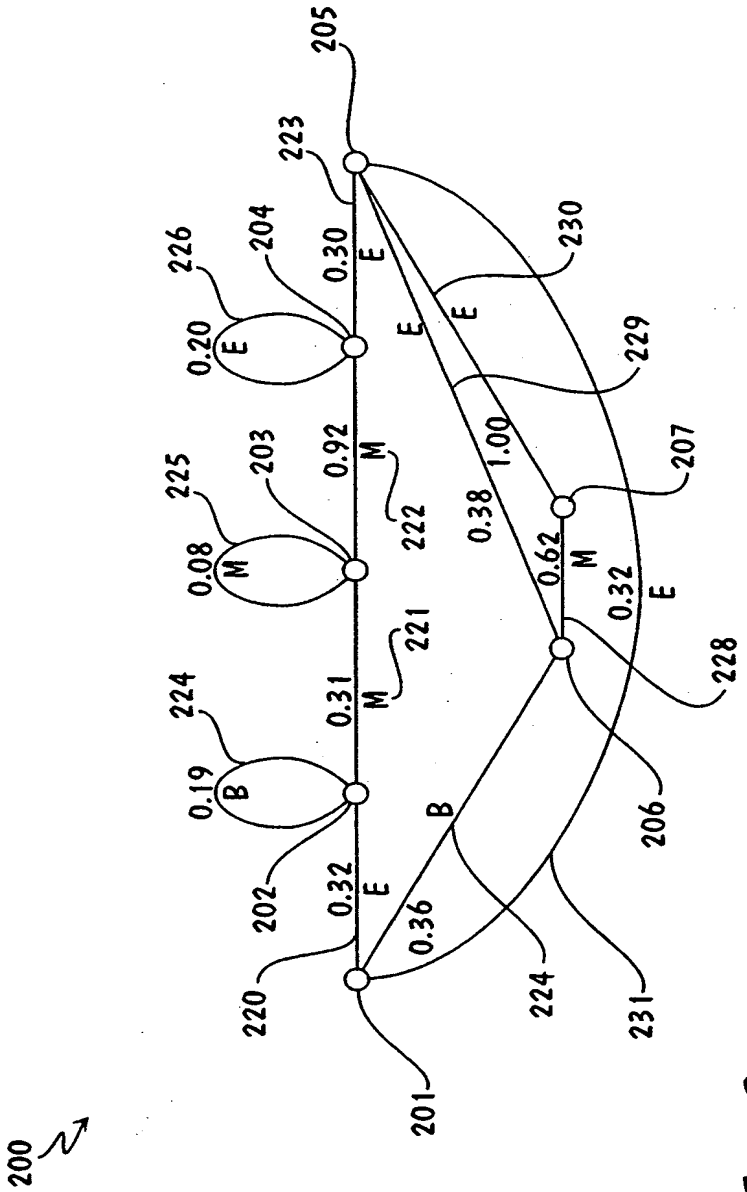


FIG. 2

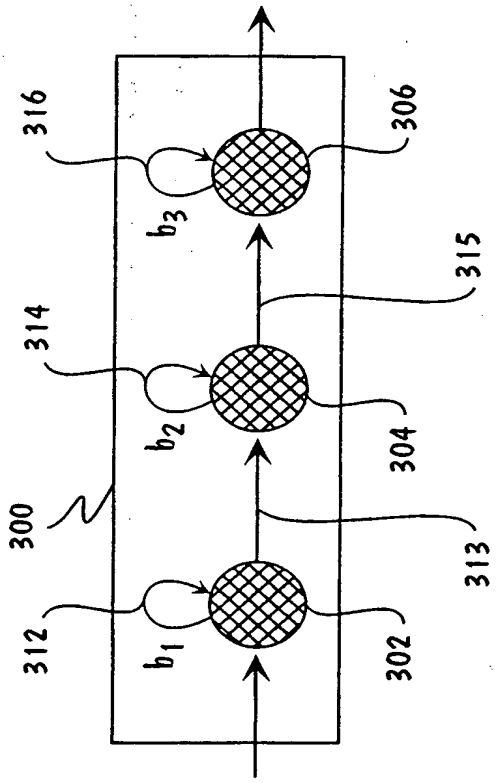


FIG. 3

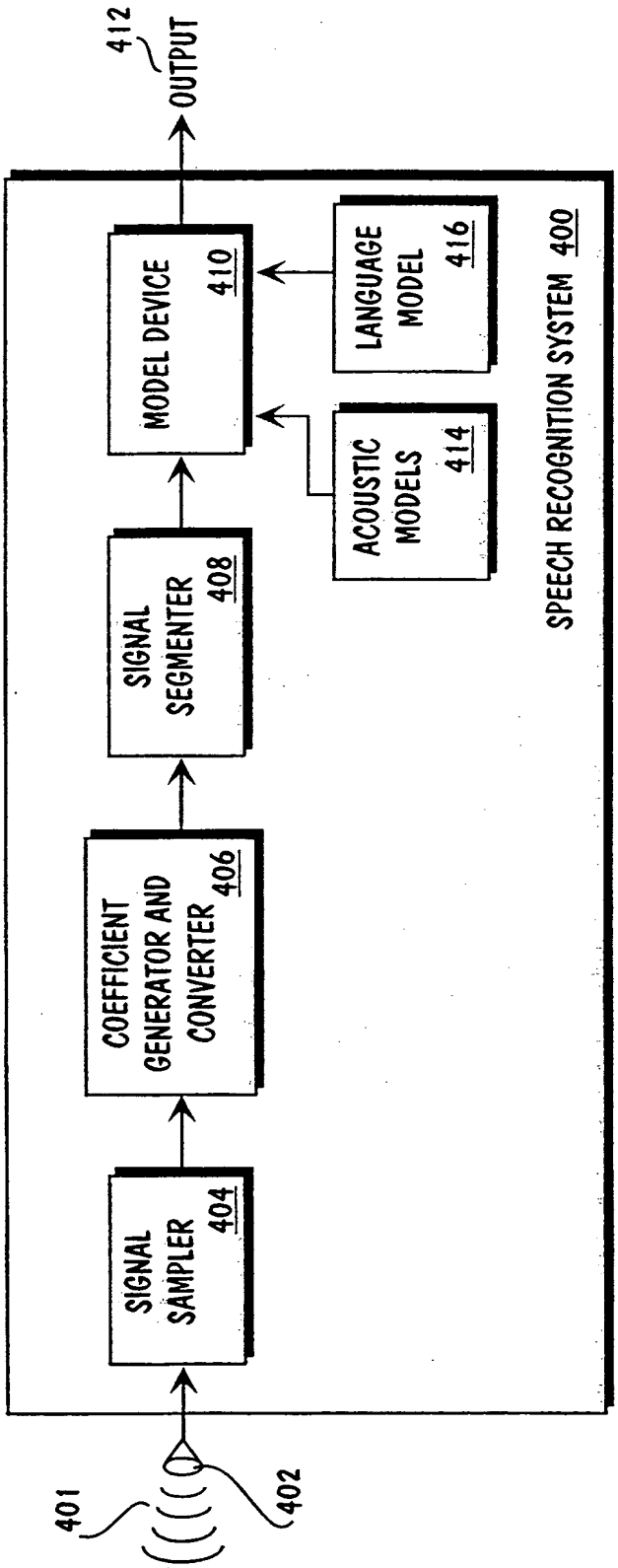


FIG. 4

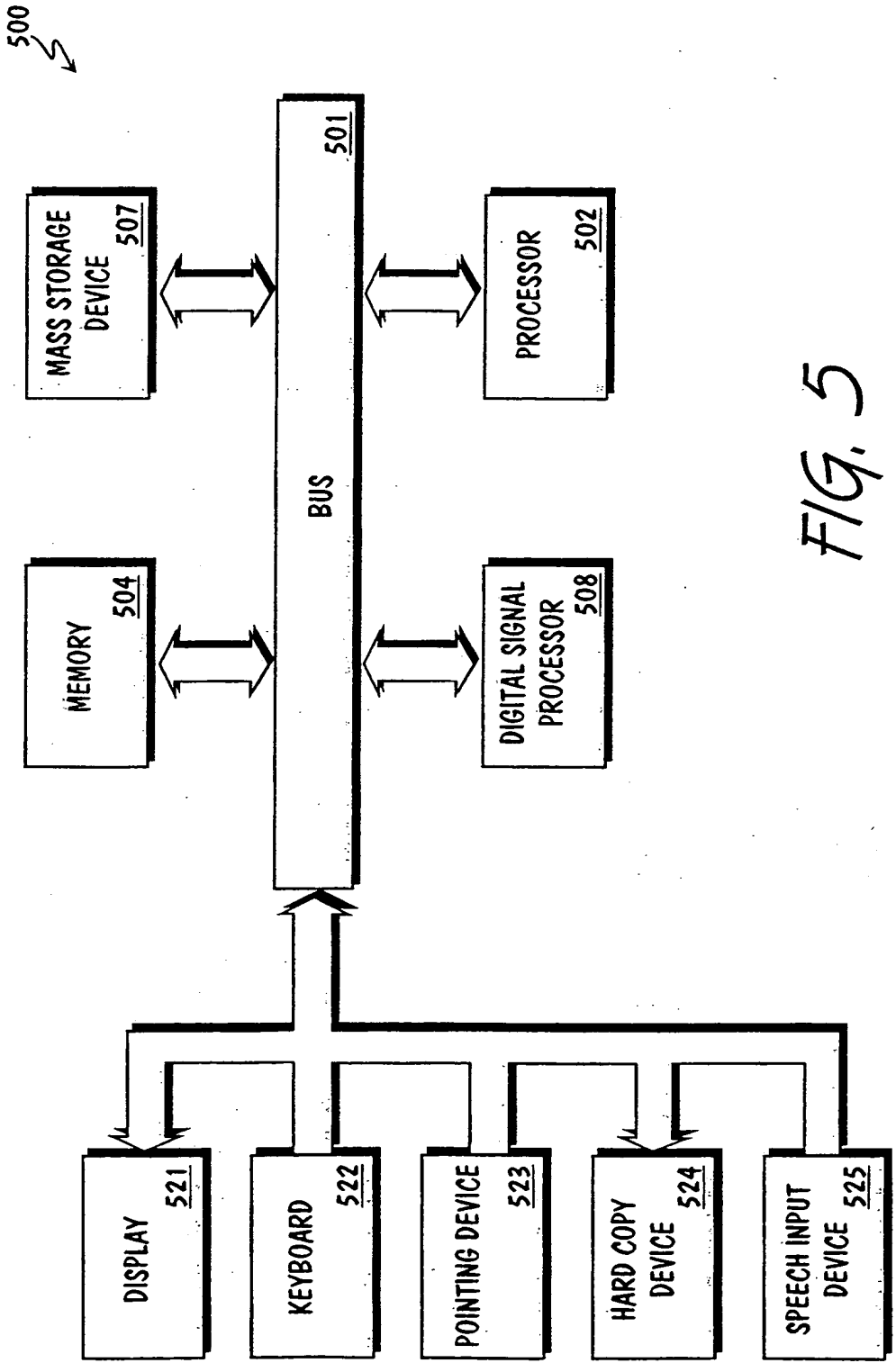

500
BUS
501
MASS STORAGE DEVICE
507
PROCESSOR
502
MEMORY
504
DIGITAL SIGNAL PROCESSOR
508
DISPLAY
521
KEYBOARD
522
POINTING DEVICE
523
HARD COPY DEVICE
524
SPEECH INPUT DEVICE
525

FIG. 5

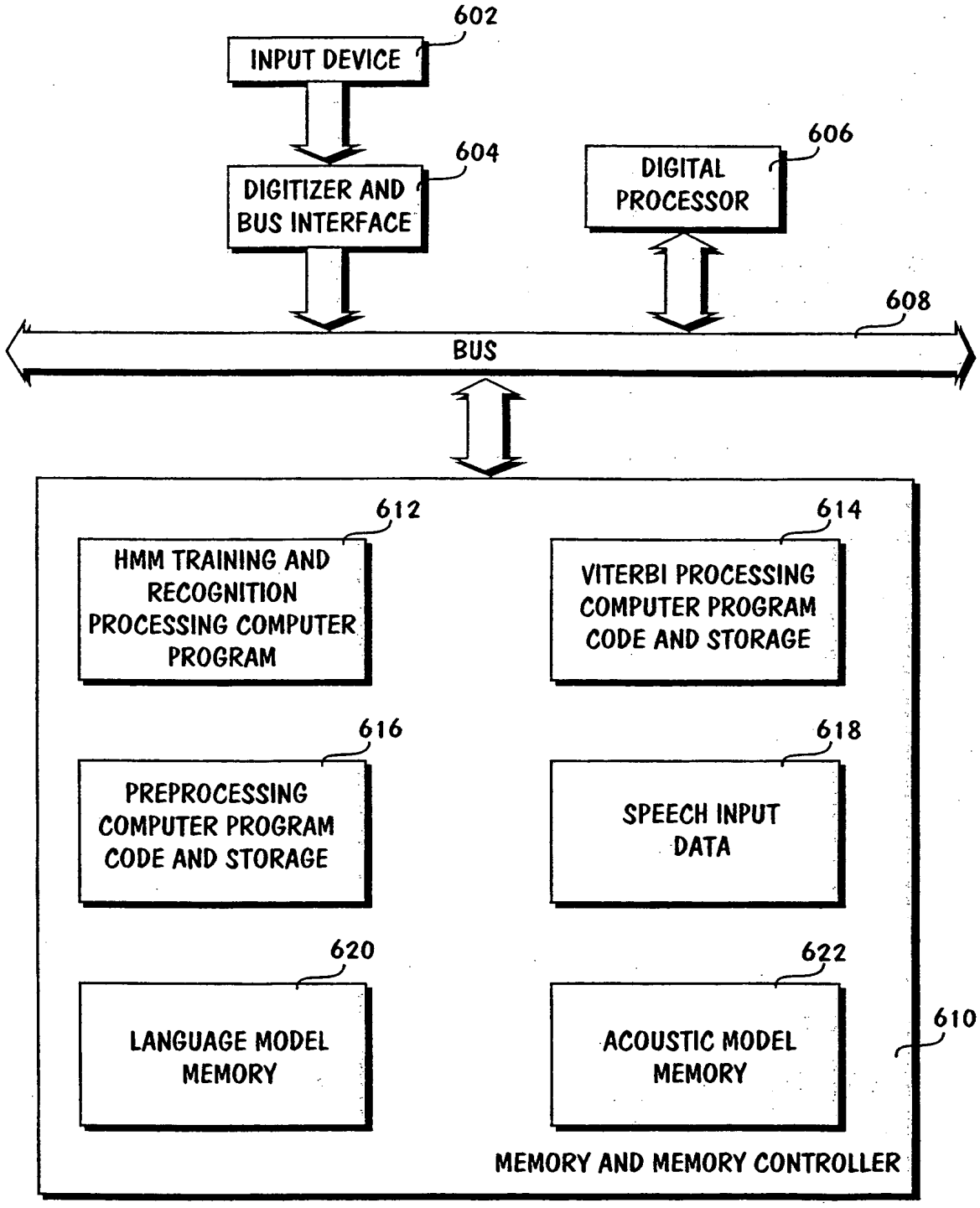


FIG. 6

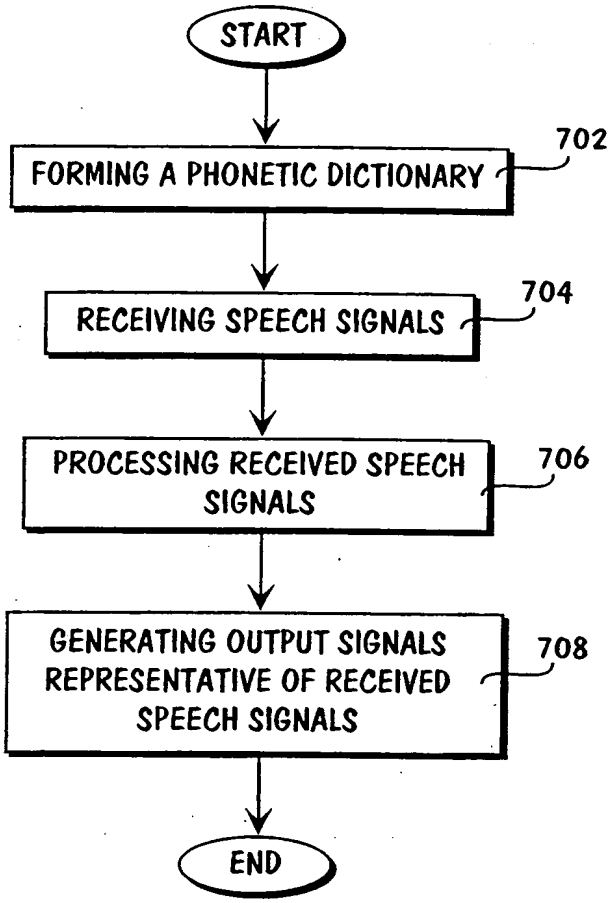


FIG. 7

P1	P2	P3	P4	P5	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
aa					iy					uw	uw	uw	uw	uw
ae					jh					v	v	v	v	v
ah					k					w	w	w	w	w
ao					l					y	y	y	y	y
aw					m					z	z	z	z	z
ay					n					zh	zh	zh	zh	zh
b					ng						ax	ax	ax	ax
ch					ow							clg	clg	cl
d					oy								ix	ix
dh					p									q
eh					r									vcl
er					s									epi
ey					sh									el
f					sil									en
g					t									dx
hh					th									
ih					uh									

FIG. 8

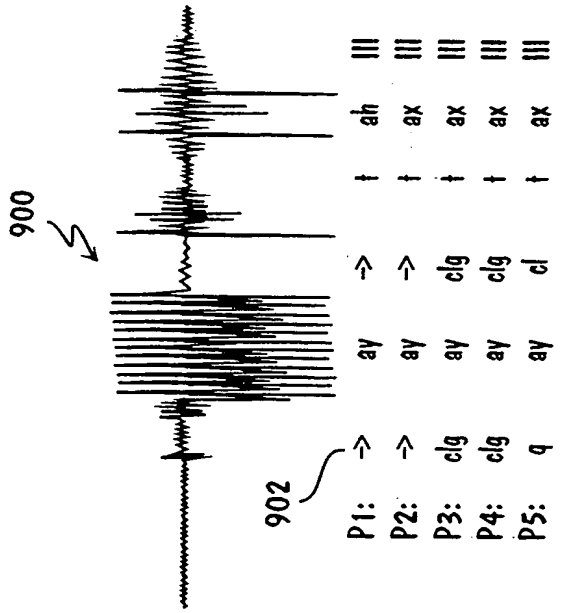

900
902
P1: -> -> ay -> t ah III
P2: -> -> ay -> t ax III
P3: clg clg ay clg t ax III
P4: clg clg ay clg t ax III
P5: q ay cl t ax III

FIG. 9

CONTEXT	DICTIONARY	PHONE SET				
		P1	P2	P3	P4	P5
MONO	CMU 110K	54.9%	55.0%	56.8%	57.3%	64.3%
TRIPHONE	SONY 5K	76.7%	88.2%	92.8%	92.2%	92.3%
	CMU 110K	67.0%	87.0%	86.0%	90.9%	90.2%

FIG. 10

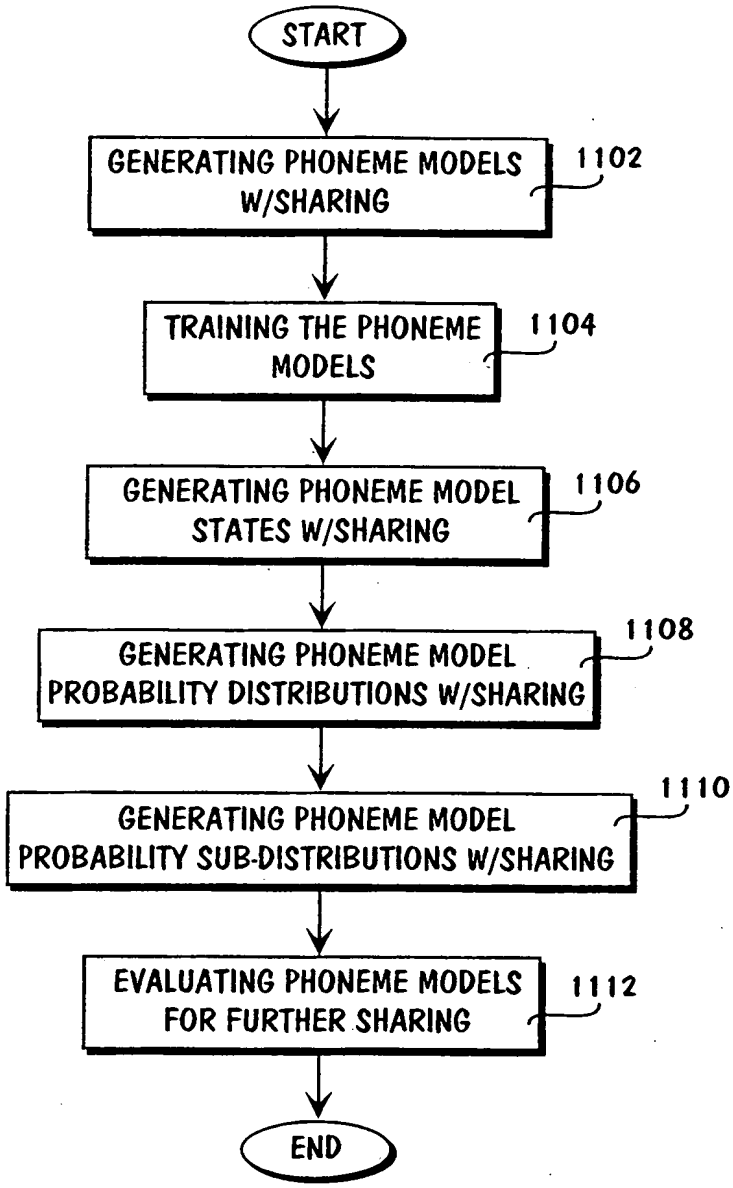

START
GENERATING PHONEME MODELS W/SHARING
1102
TRAINING THE PHONEME MODELS
1104
GENERATING PHONEME MODEL STATES W/SHARING
1106
GENERATING PHONEME MODEL PROBABILITY DISTRIBUTIONS W/SHARING
1108
GENERATING PHONEME MODEL PROBABILITY SUB-DISTRIBUTIONS W/SHARING
1110
EVALUATING PHONEME MODELS FOR FURTHER SHARING
1112
END

FIG. 11

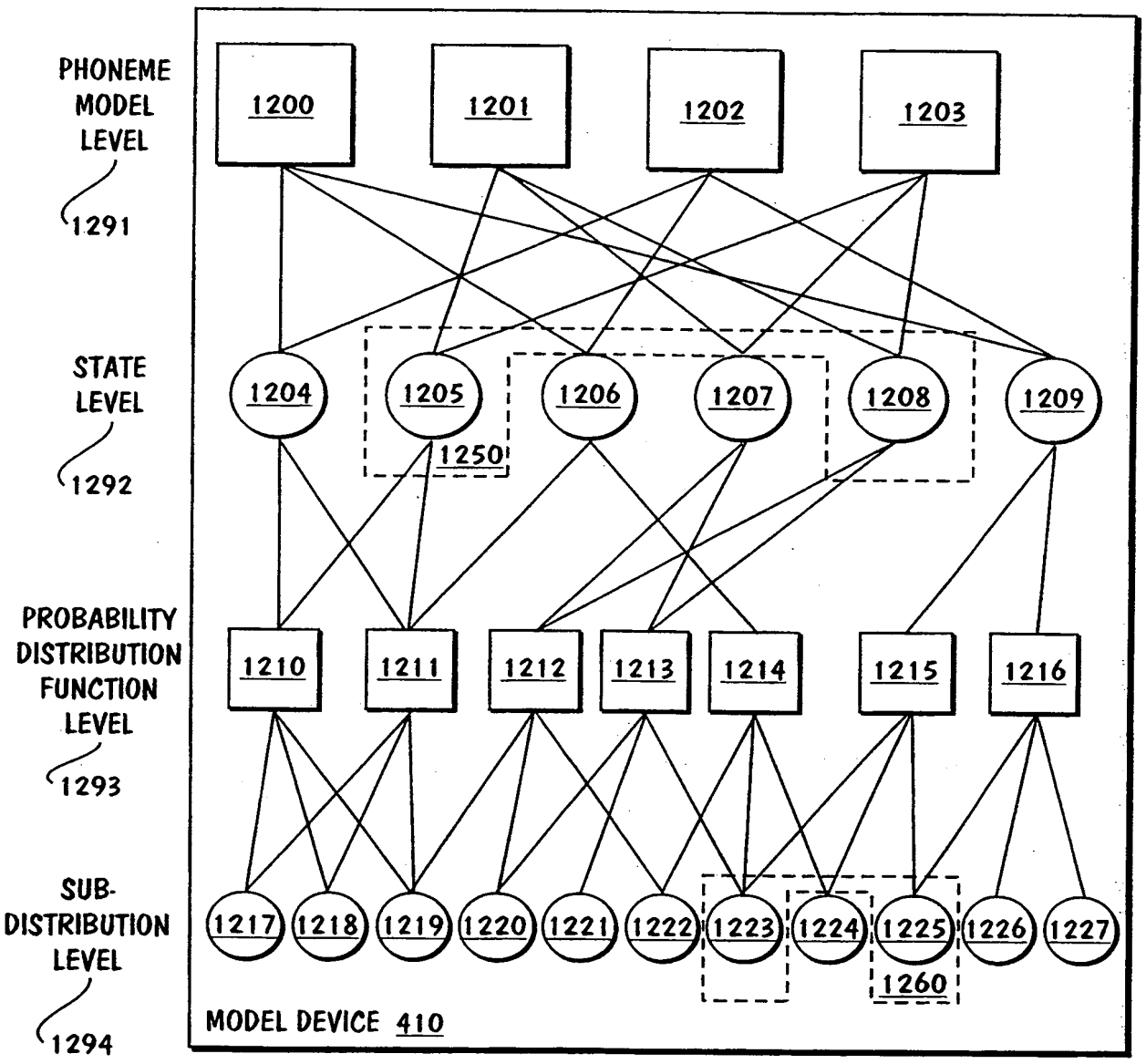

PHONEME MODEL LEVEL
1291
STATE LEVEL
1292
PROBABILITY DISTRIBUTION FUNCTION LEVEL
1293
SUB-DISTRIBUTION LEVEL
1294
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1250
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1260
MODEL DEVICE 410

FIG. 12

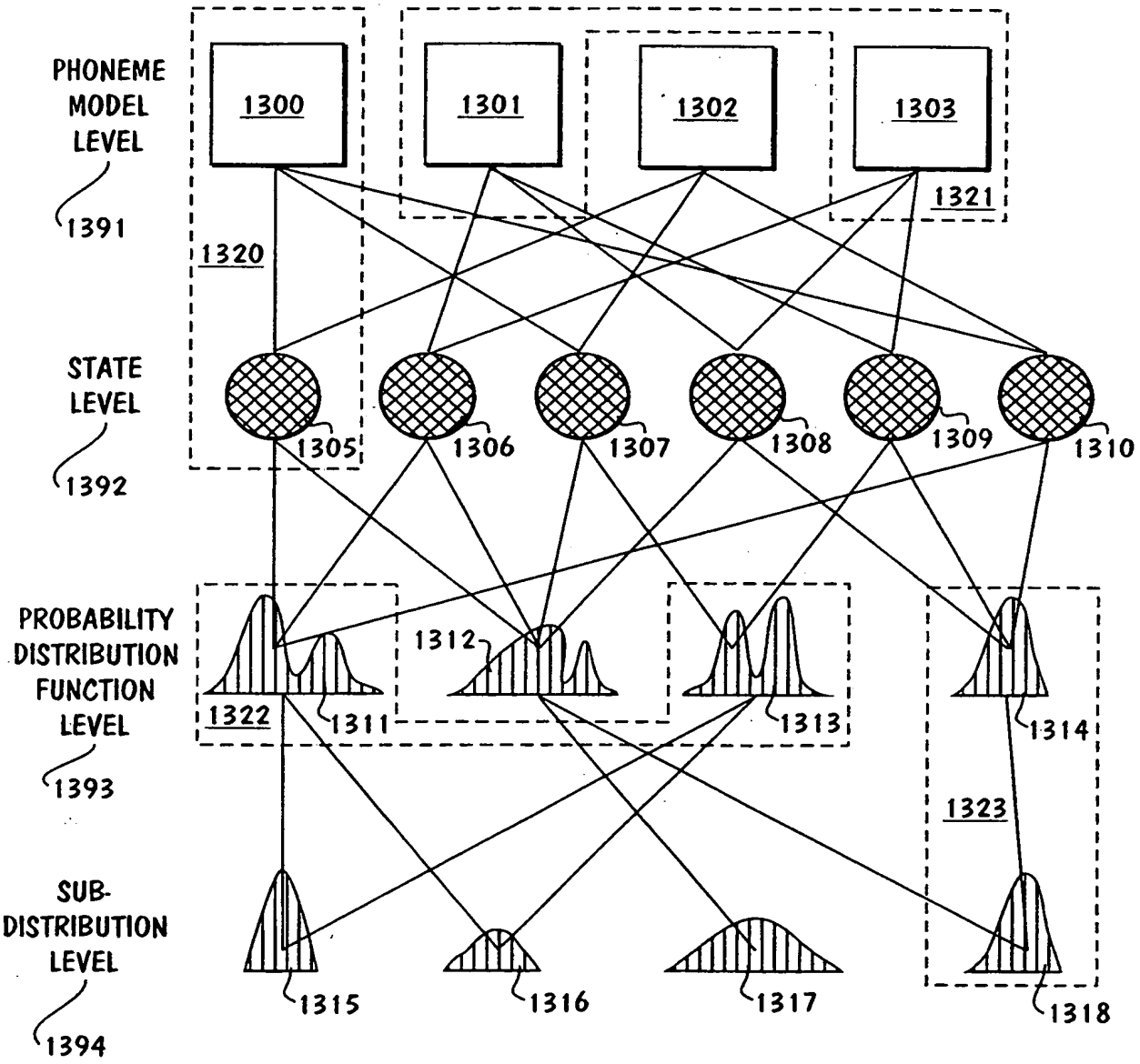

PHONEME MODEL LEVEL
1391
STATE LEVEL
1392
PROBABILITY DISTRIBUTION FUNCTION LEVEL
1393
SUB-DISTRIBUTION LEVEL
1394
1300
1301
1302
1303
1320
1321
1305
1306
1307
1308
1309
1310
1312
1322
1311
1313
1314
1323
1315
1316
1317
1318

FIG. 13

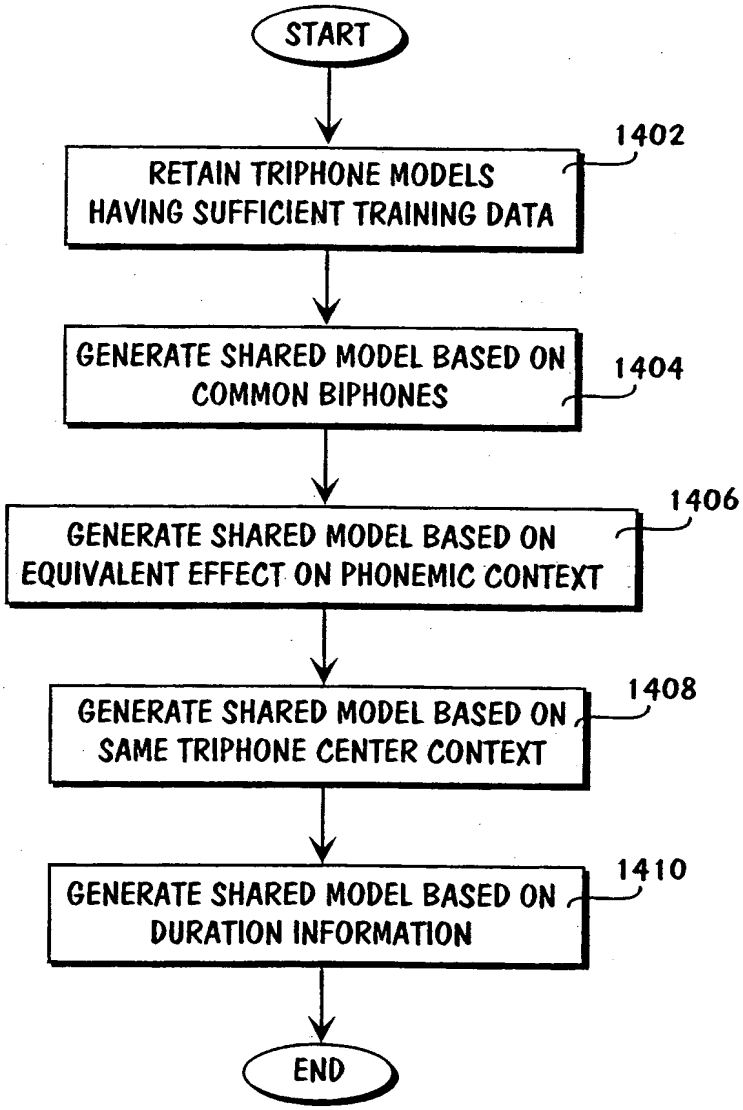

START
RETAIN TRIPHONE MODELS HAVING SUFFICIENT TRAINING DATA
1402
GENERATE SHARED MODEL BASED ON COMMON BIPHONES
1404
GENERATE SHARED MODEL BASED ON EQUIVALENT EFFECT ON PHONEMIC CONTEXT
1406
GENERATE SHARED MODEL BASED ON SAME TRIPHONE CENTER CONTEXT
1408
GENERATE SHARED MODEL BASED ON DURATION INFORMATION
1410
END

FIG. 14

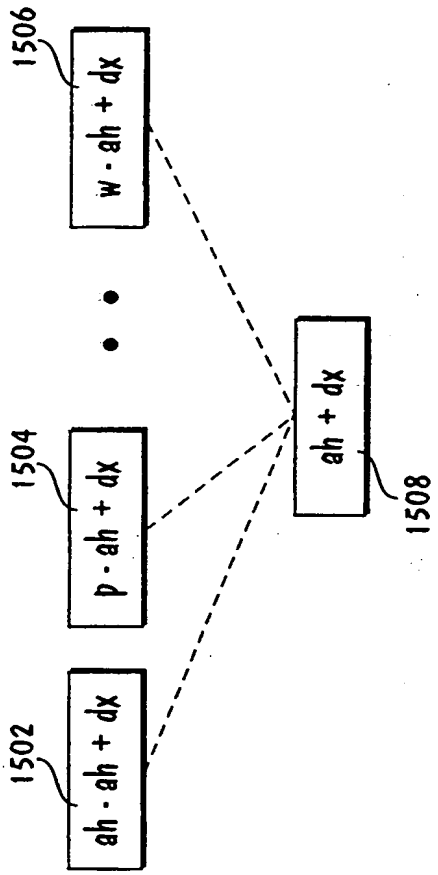


FIG. 15

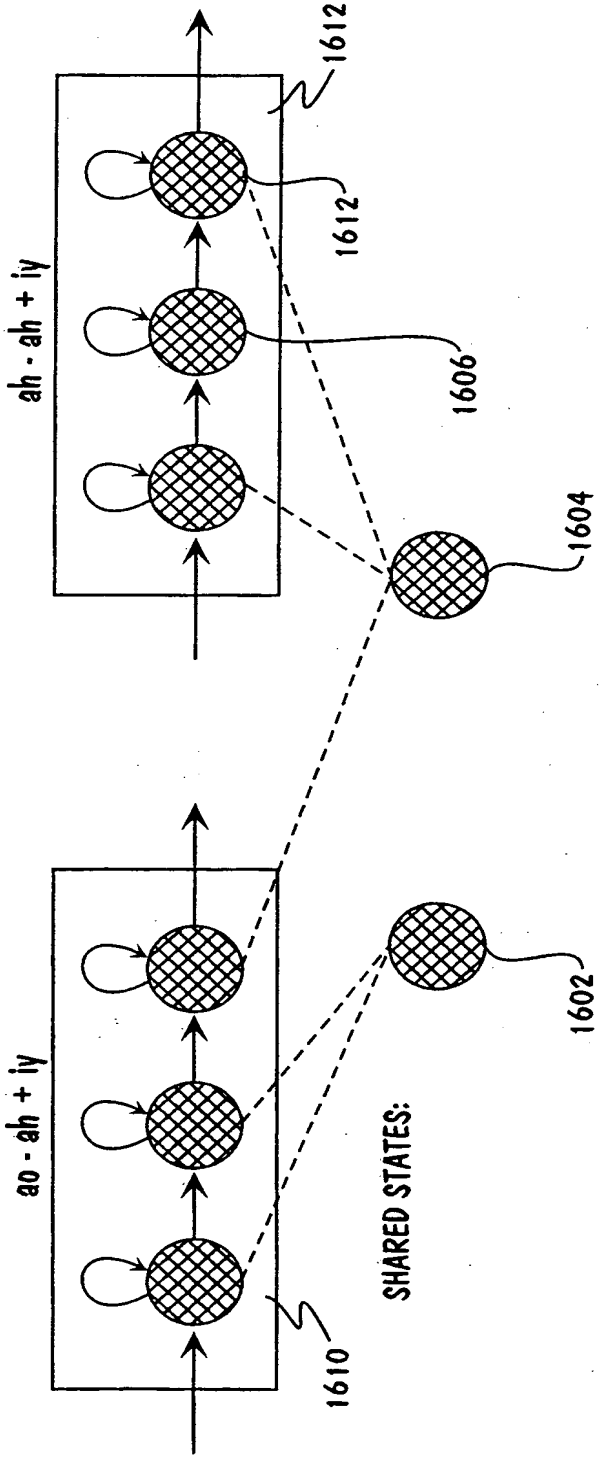


FIG. 16

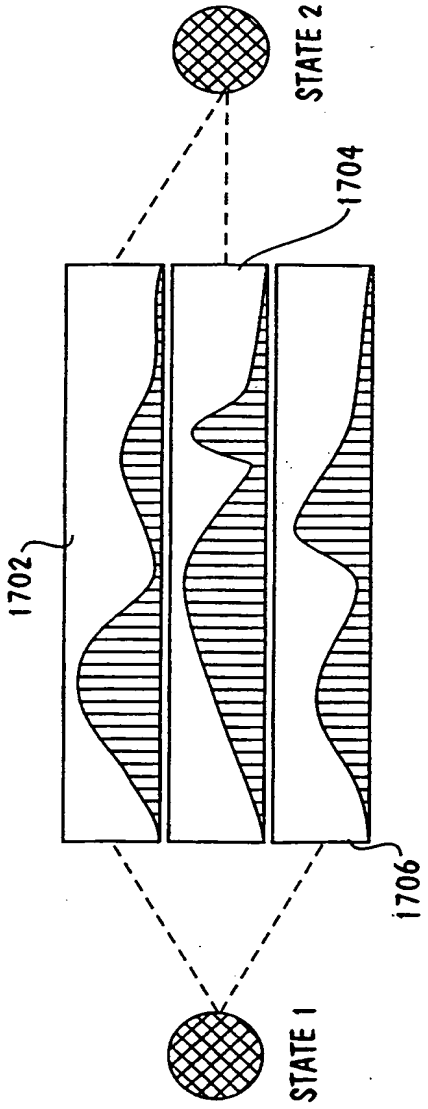


FIG. 17

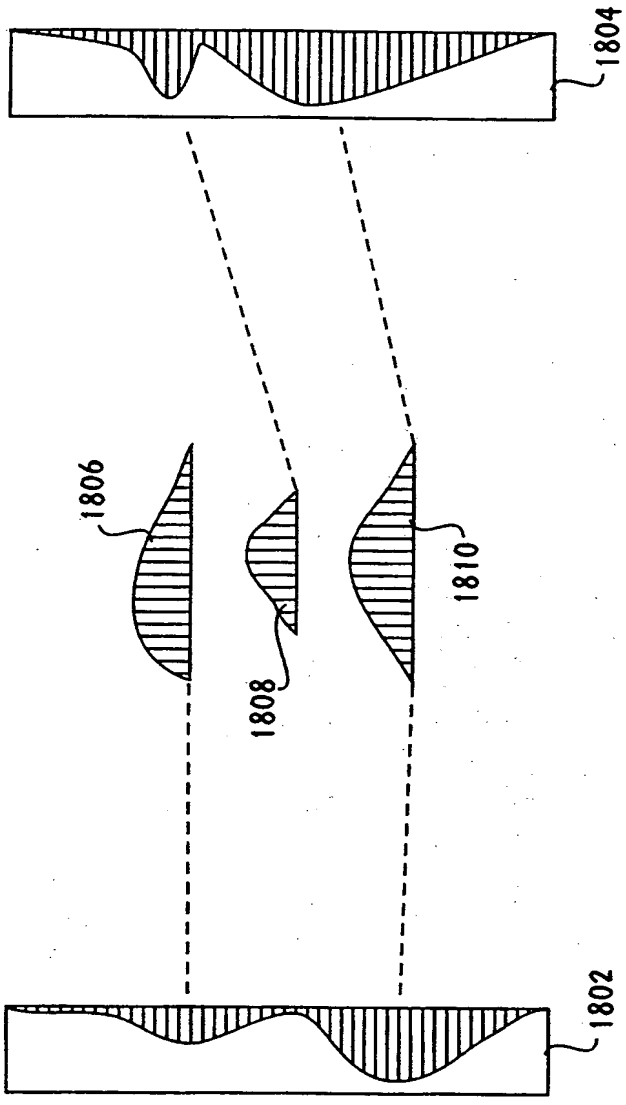


FIG. 18

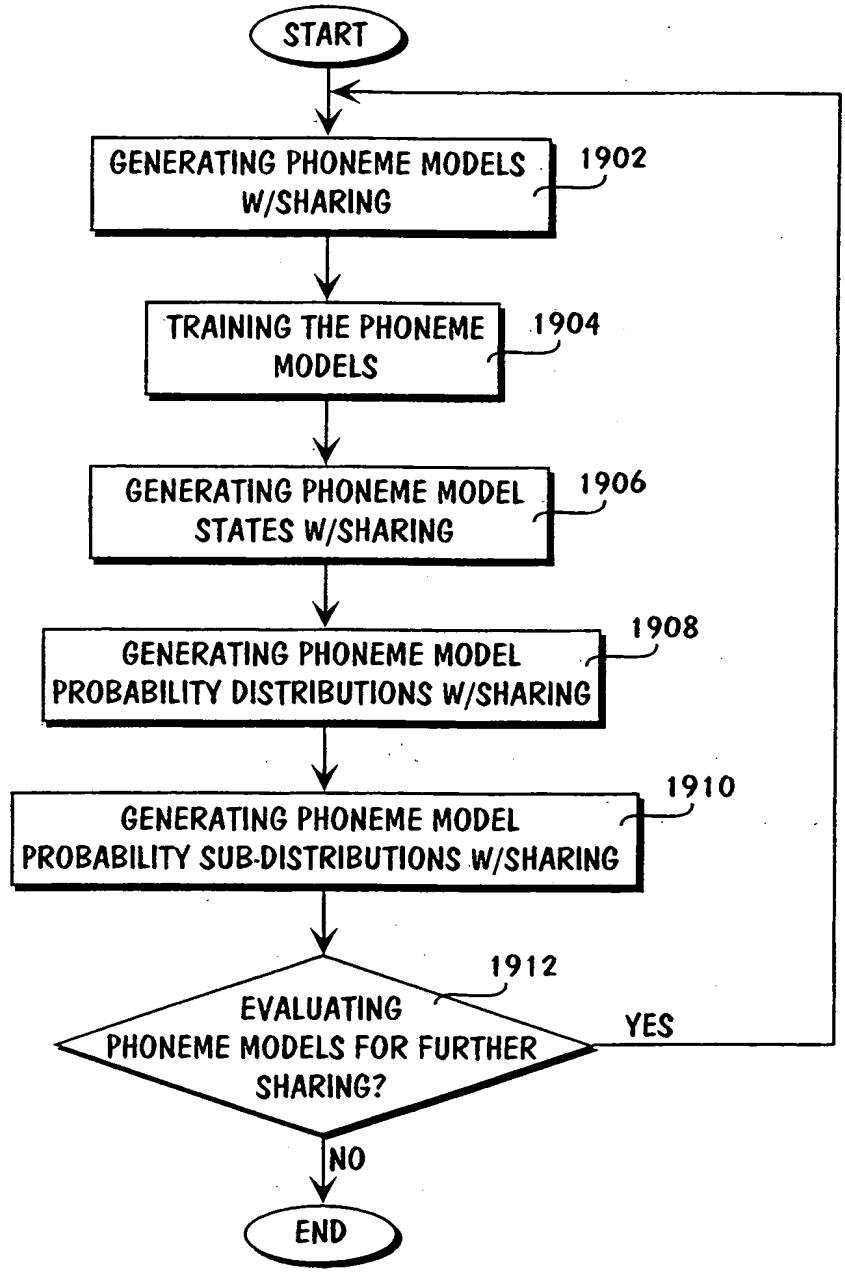

START
GENERATING PHONEME MODELS W/SHARING
1902
TRAINING THE PHONEME MODELS
1904
GENERATING PHONEME MODEL STATES W/SHARING
1906
GENERATING PHONEME MODEL PROBABILITY DISTRIBUTIONS W/SHARING
1908
GENERATING PHONEME MODEL PROBABILITY SUB-DISTRIBUTIONS W/SHARING
1910
EVALUATING PHONEME MODELS FOR FURTHER SHARING?
1912
YES
NO
END

FIG. 19

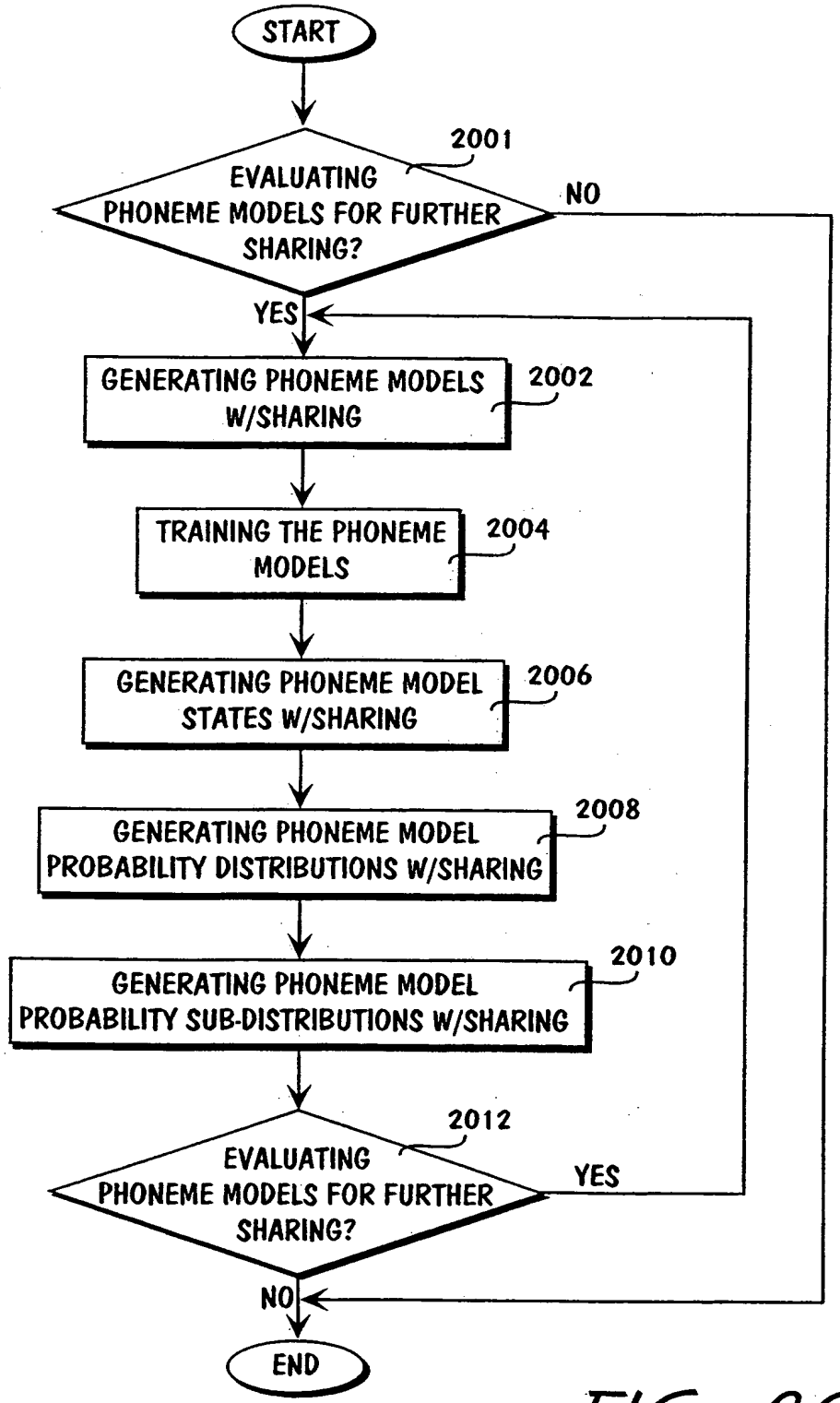


FIG. 20

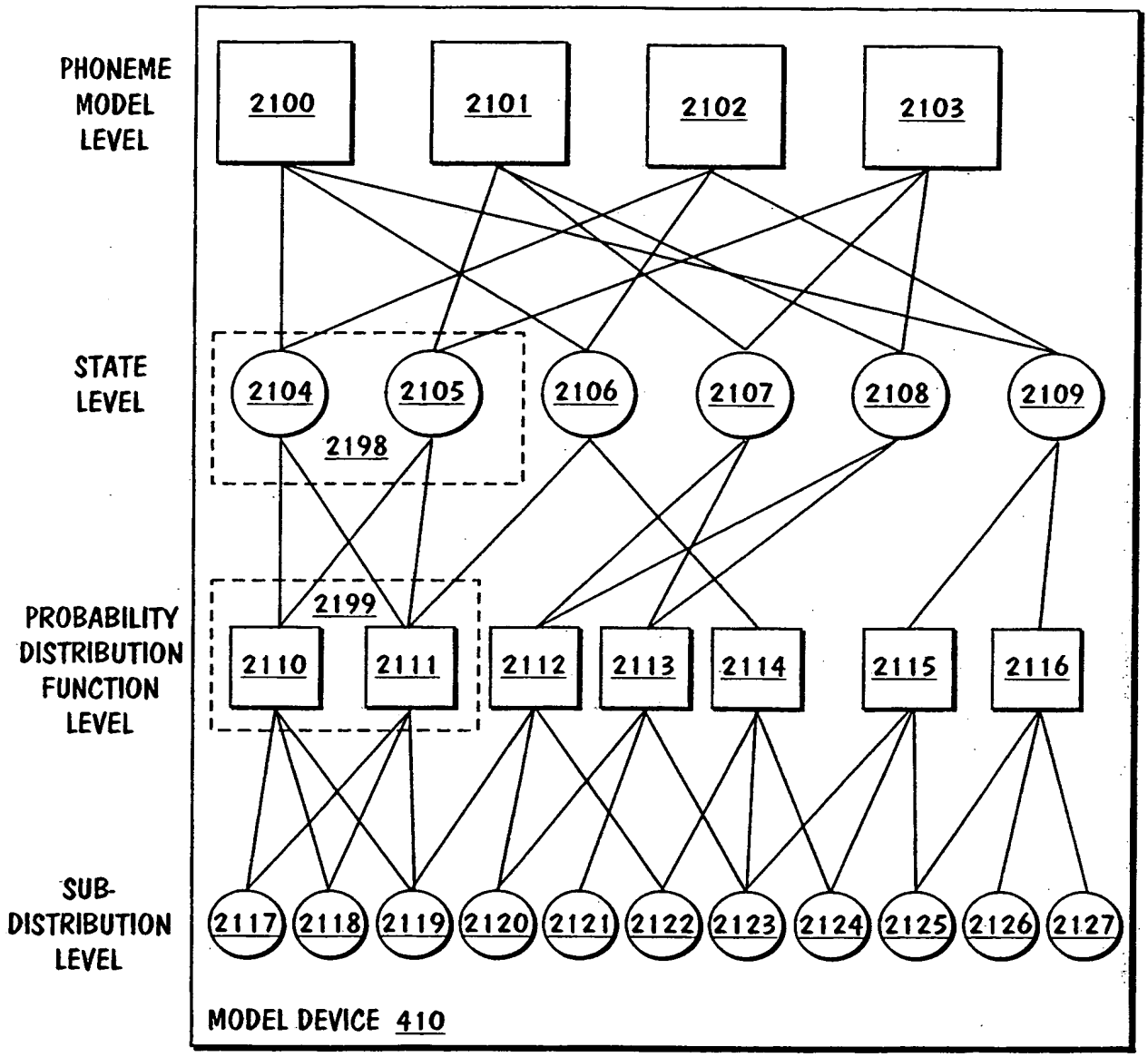

PHONEME MODEL LEVEL
STATE LEVEL
PROBABILITY DISTRIBUTION FUNCTION LEVEL
SUB-DISTRIBUTION LEVEL
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2198
2199
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
MODEL DEVICE 410

FIG. 21

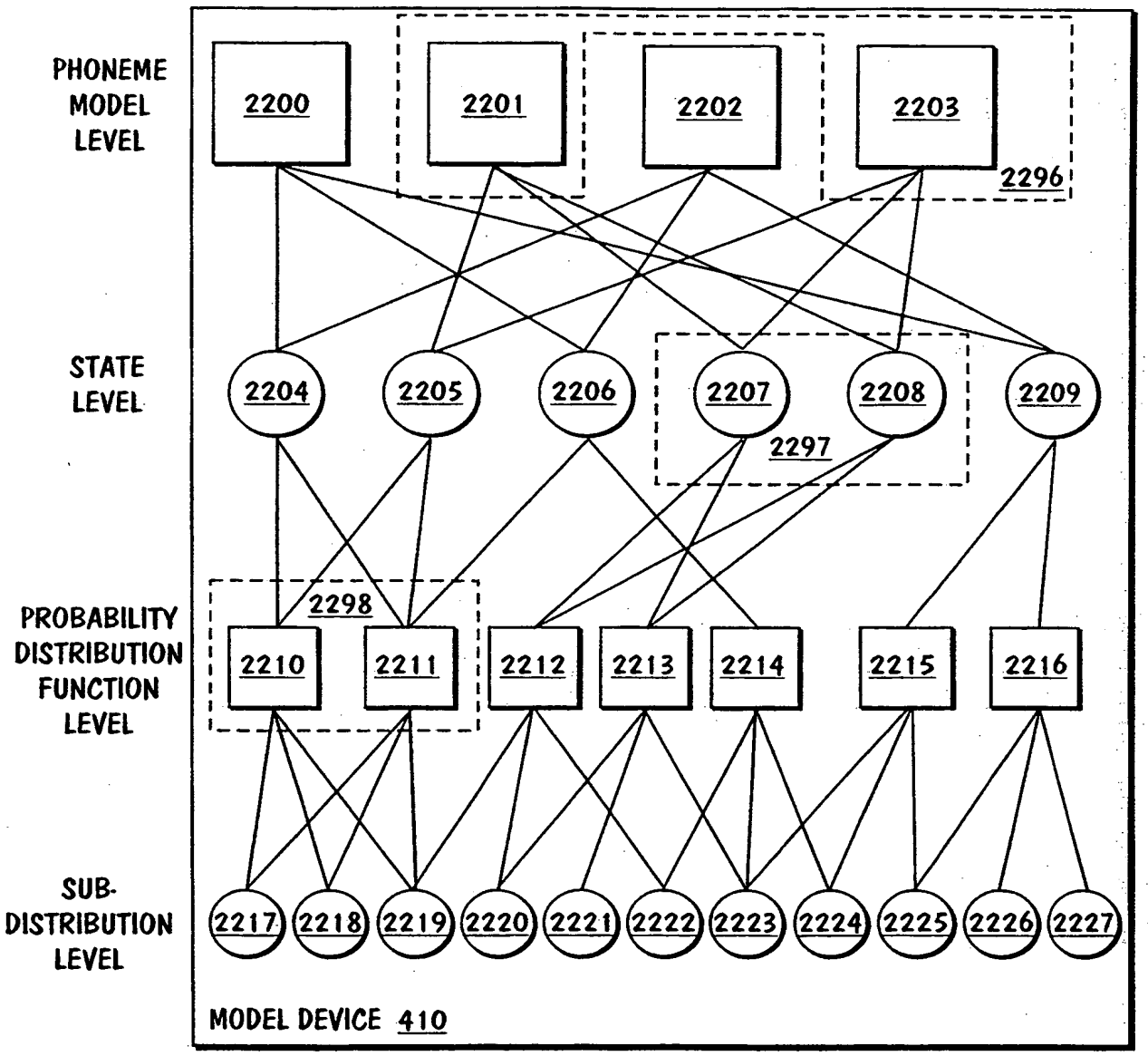

PHONEME MODEL LEVEL
STATE LEVEL
PROBABILITY DISTRIBUTION FUNCTION LEVEL
SUB-DISTRIBUTION LEVEL
2200
2201
2202
2203
2296
2204
2205
2206
2207
2208
2209
2297
2298
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
MODEL DEVICE 410

FIG. 22

HMMs	MONO-PHONE M0	LEFT BIPHONE M1	LEFT BIPHONE M2	RIGHT BIPHONE M3	8114 STATES M4	2583 STATES M5
MIXTURE	16	4	16	16	16	16
TOP 1	36.9%	34.5%	51.4%	64.7%	67.2%	66.4%
TOP 2	47.7%	42.4%	60.2%	77.7%	76.2%	76.6%
TOP 3	53.4%	45.7%	63.6%	82.6%	79.6%	80.7%
# OF STATES	147	7350	7350	7350	8114	2583
REC-SPEED	1.9	4.0	9.2	8.7	10.6	9.2

FIG. 23

DIFFERENT DICTIONARIES (CMU)	CMU (ORIGINAL) D1	OPTIONAL CLOSURE D2	MULTIPLE SPELLINGS D3	INCLUDE REAL DATA D4
TOP 1	44.9%	66.4%	79.2%	83.9%
TOP 2	53.5%	76.6%	87.4%	90.9%
TOP 3	57.7%	80.7%	90.2%	93.3%

FIG. 24

DIFFERENT DICTIONARIES (CMU)	CMU (ORIGINAL) D1	OPTIONAL CLOSURE D2	MULTIPLE SPELLINGS D3	INCLUDE REAL DATA D4
TOP 1	48.8%	72.4%	81.4%	84.8%
TOP 2	57.3%	82.4%	89.4%	91.8%
TOP 3	61.2%	86.1%	92.1%	94.2%

FIG. 25

DIFFERENT DICTIONARIES (P5 PHONES)	110K CMU D3	50K SONY D5	5K MOST COMMON D6	5K US CITIES D7
TOP 1	79.2%	79.6%	85.9%	87.0%
TOP 2	87.4%	88.6%	91.2%	91.1%
TOP 3	90.2%	91.4%	92.3%	92.6%

FIG. 26

DIFFERENT DICTIONARIES (P5 PHONES)	110K CMU D3	50K SONY D5	5K MOST COMMON D6	5K US CITIES D7
TOP 1	83.9%	85.7%	90.1%	90.0%
TOP 2	90.9%	92.8%	95.2%	93.4%
TOP 3	93.3%	94.9%	96.4%	94.4%
REC-SPEED	10.1	11.9	11.9	9.1

FIG. 27

DIFFERENT DICTIONARIES (CMU)	CMU (ORIGINAL) D1	OPTIONAL CLOSURE D2	MULTIPLE SPELLINGS D3	INCLUDE REAL DATA D4
TOP 1	84.8%	85.9%	89.9%	90.3%
TOP 2	91.8%	93.1%	95.2%	93.5%
TOP 3	94.2%	95.3%	96.8%	94.6%
REC-SPEED	12.0	12.1	11.6	8.0

FIG. 28

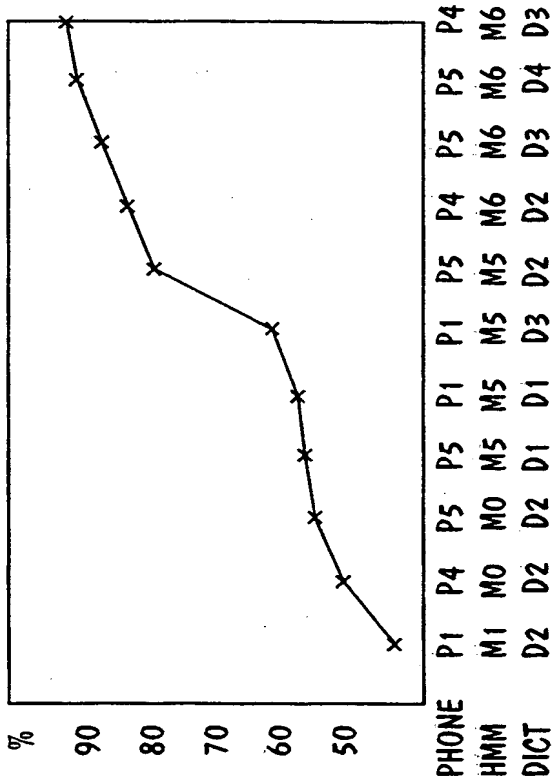

%
90
80
70
60
50
PHONE P1 P4 P5 P5 P1 P5 P1 P5 P4 P5 P5 P4
HMM M1 M0 M0 M5 M5 M5 M5 M6 M6 M6 M6 M6
DICT D2 D2 D2 D1 D1 D3 D2 D2 D3 D4 D3

FIG. 29

[13] N. Iwahashi, et al(1998). Stochastic Features For Noise Robust Speech Recognition. ICASSP'98, v. 2 pp. 633.

[14] Shan Zhu, et al(1996). Feature Parameter Curve Method For High Performance NN-Based Speech Recognition. ICASSP'96, v. 1 pp. 1.

[15] Yih-Ru Wang, et al(1998). Mandarin Telephone Speech Recognition For for Automatic Telephone Number Directory Service. ICASSP'1998, v. 2 pp. 841.

[16] Joachim Kohlen(1998). Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks. ICASSP'98, v. 2 pp. 417.

[17] Soren K. Riis(1998). Hidden Neural Networks: Application To Speech Recognition. ICASSP'98, v. I pp. 1117.

[18] Yukhyun Chang, et al(1998). Improved Model Parameter Compensation Methods For Noise-Robust Speech Recognition. ICASSP'98, v. 1 pp. 561.

[19] Tony Robinson and James Christie (1998). Time-First Search For Large Vocabulary Speech Recognition. ICASSP'98, v. 2 pp. 829.

[20] Merhyar Mohri, et al(1998). Full Expansion of Context-Dependent Networks In Large Vocabulary Speech Recognition. ICASSP'98, v. 2 pp. 665.

[21] L.R. Bahl, et al(1995). Performance Of The IBM Large Vocabulary Continuous Speech Recognition System On the ARPA Wall Street Journal Task. ICASSP'95, v. 1 pp. 41.

[22] P. Laderfoged (1993), A Course in Phonetics. Harcourt Brace Jovanovich College Publishers, New York, 3rd Edition.

FIG. 30a

DICT. SIZE	LANGUAGE	C/I	CONDITION	REFERENCE	TOP1 [TOP3]
110K	ENGLISH	I	CLEAN	SONY, USA	84.8 [94.2]
5000	ENGLISH	I	CLEAN	SONY, USA	89.9 [96.8]
5075	JAPANESE	I	CAR NOISE SNR = 4.3	SONY, JAPAN [13]	90.0
1345	MANDARIN	I	CLEAN SYLLABLES 1 SPEAKER	NANYAN TECH. U SINGAPORE[14]	88.2 [98.9]
1922	MANDARIN	I	TELEPHONE BLANK NAMES	CHAIO TUNG UNIV., ROC[15]	94.8 [97.9]
685	ENGLISH	I	TELEPHONE	SIEMENS, GERMAN[16]	64.5
600	ENGLISH	I	TELEPHONE	TECH. UNIV, DENMARK[17]	85.8
75	KOREAN	I	CLEAN	LGIC, KOREAN [18]	93.9
65K	ENGLISH	C	CLEAN	CAMBRIDGE, UK[19]	83.2
10K	ENGLISH	C	CLEAN NAB	AT&T, USA[20]	~82
160K	ENGLISH	C	CLEAN NAB	AT&T, USA[20]	~74
60K	ENGLISH	C	CLEAN WSJ	IBM, USA[21]	90.6

C: CONTINUOUS
I: ISOLATED

FIG. 30b

INTERNATIONAL SEARCH REPORT

Inter. nal Application No
PCT/US 99/25978

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/02 G10L15/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 799 276 A (MALKOVSKY MIKHAIL ET AL) 25 August 1998 (1998-08-25) column 5, line 33 - line 38 column 12, line 66 -column 13, line 29 column 38, line 24 - line 31 column 38, line 57 -column 39, line 9 column 41, line 13 - line 28 column 41, line 48 -column 42, line 13	1,3,4,9, 10,23, 25,35, 37,48
X	US 5 581 655 A (COHEN MICHAEL H ET AL) 3 December 1996 (1996-12-03) column 2, line 24 -column 3, line 56 column 10, line 61 -column 11, line 14 column 6, line 47 - line 53	1,23,35, 48

-/—

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 March 2000	08/03/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Ramos Sánchez, U

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter. nal Application No
PCT/US 99/25978

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 750 293 A (CANON KK) 27 December 1996 (1996-12-27) page 2, line 7 - line 26	2,24,42, 56
A	US 5 715 367 A (GILLICK LAURENCE S ET AL) 3 February 1998 (1998-02-03) column 4, line 14 - line 16 column 4, line 24 - line 28 column 4, line 60 -column 5, line 5 column 5, line 25 - line 33	1,2,23, 24,35, 42,48,56
A	US 5 625 749 A (GLASS JAMES R ET AL) 29 April 1997 (1997-04-29) column 18, line 1 - line 13	1,23,48
A	TAKAHASHI S ET AL: "FOUR-LEVEL TIED-STRUCTURE FOR EFFICIENT REPRESENTATION OF ACOUSTIC MODELING" PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP '95), DETROIT, MI, USA, 9 - 12 May 1995, pages 520-523, XP000658045 IEEE, NEW YORK, US, ISBN: 0-7803-2432-3 the whole document	1-58
A	US 5 502 790 A (YI JIE) 26 March 1996 (1996-03-26) column 2, line 5 - line 29	1-58

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Intern. nal Application No
PCT/US 99/25978

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5799276 A	25-08-1998	NONE	
US 5581655 A	03-12-1996	US 5268990 A	07-12-1993
		CA 2099978 A	01-08-1992
		EP 0573553 A	15-12-1993
		JP 6505349 T	16-06-1994
		WO 9214237 A	20-08-1992
EP 0750293 A	27-12-1996	JP 9006386 A	10-01-1997
		US 5812975 A	22-09-1998
US 5715367 A	03-02-1998	NONE	
US 5625749 A	29-04-1997	NONE	
US 5502790 A	26-03-1996	JP 5173588 A	13-07-1993
		JP 5188989 A	30-07-1993
		JP 5188990 A	30-07-1993

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.